Dipartimento di Scienze dell'Ambiente Forestale e delle sue Risorse

Dottorato di Ricerca in Scienze e Tecnologie per la Gestione Forestale e Ambientale

XVIII Ciclo

ESTIMATING AND MAPPING FOREST INVENTORY VARIABLES USING THE K-NN METHOD: MOCUBA DISTRICT CASE STUDY - MOZAMBIQUE

AGR/05 –Assestamento Forestale e Selvicoltura

Coordinatore: Prof. G. Piovesan

Tutor: Prof. P. Corona

Dottorando: Carla R. Pereira

January, 2006

Table of contents

Acknowledgements

This study was supported by the National Forest Inventory Unit within the Ministry of Agriculture of Mozambique which kindly provided the data used. To my colleagues from the Forest Inventory Unit I would like to express my special gratitude.

I thank my supervisor Prof. Piermaria Corona for his guidance and I would like to extend my immense gratitude to Dr. Paolo Calvani and Dr. Davide Travaglini for their patience and support with the *k*-NN and GIS softwares.

Abstract

Forest inventory covering vast areas provides a high number of variables and information that need to be processed from plot level to the overall inventoried area. The increasing availability of remotely sensed data has made possible to obtain a synoptic view over large areas, although with less accuracy than the field measurements. The introduction of georeferenced sample plots in forest inventories created the possibility of combining field data to satellite images, that is multisource forest inventories, and to expand the information from plot level to the overall image.

In this study, forest variables were estimated by means of Landsat ETM images and existing forest inventory field plots data. The study was carried out for Mocuba district in Zambézia province – Mozambique, characterized by typical african miombo forests with predominance of *Julbernardia globiflora* and *Brachystegia sp.*, covering an area of about 500 000 hectares. The non-parameteric *k*-nearest neighbour (*k*-NN) estimator was used for estimating and mapping of three selected forest variables: total volume, commercial volume and forest density. The method has been applied successfully in temperate and boreal forests but only a few studies regarding species richness prediction can be found in the literature for tropical and sub-tropical regions.

To estimate the selected forest variables, *k*-NN parameters were firstly calibrated to the district data in order to determine the most appropriated distance type and number of neighbours. Results obtained are encouraging and in line with those mentioned in several literature and studies, mostly for Nordic countries. When the *k*-NN estimator and the central point/pixel value was used to estimate forest variables high relative errors were obtained, from 76 to 99%. When the average values of the pixels within sample plots with 30 meters buffer area was employed to estimate the forest variables the relative error dropped to 47 - 63%. Errors obtained are far higher than those allowed for forest inventory (usually less than 20%) and are not accurate enough for forest management purposes at district and concessions levels. However, the mapping of forest variables with a known error is a powerful instrument for strategic decision making. Forest variables estimates derived by the method can easily produced and used in areas where no other information exists or for non-sampled areas.

Keywords: *Remote sensing*; *k-NN; multisource forest inventory; miombo forests.*

Abstract

Gli inventari forestale su vaste superfici forniscono un elevato numero di variabili ed informazioni che necessitano di essere processate dal livello di aree di saggio a quello della area complessivamente sottoposta all'inventario. La disponibilità crescente di immagini satellitari rende possibile ottenere una visione sinottica su ampie aree, ancorché con minor precisione rispetto alle misurazioni sul campo. L'introduzione di campioni georeferenziati negli inventari forestali ha offerto la possibilità di combinare dati di campo con gli immagini satellitari, ossia inventari forestali multisource, e di espandere le osservazione puntuale al suolo alla immagine complessiva.

Nel presente studio, le variabili forestali sono stimate per mezzo di immagini Landsat ETM e dati a terra del' inventario forestal. Lo studio è stato condotto per il distretto di Mocuba, nella provincia della Zambesia - Mozambico, caratterizzata da foreste del tipico *miombo* africano con la predominanza di *Julbernardia globiflora* e *Brachystegia sp,* che ricopre un'area di circa 500 000 ettari. Il algoritmo non parametrico *k*-vicino più vicino (*k*-NN) è stato impiegato per la valutazione e la mappatura di tre variabili forestali selezionate: il volume totale, il volume commerciale e la densità forestale. Il metodo è stato in precedenza e con successo applicato in foreste temperate e boreali ma sono disponibili solo alcuni studi relativi alla previsione di ricchezza di specie nella letteratura sulle regioni tropicali e subtropicali.

Al fine di stimare le variabili forestali selezionate, i parametri *k*-NN sono stati calcolati per i dati del distretto per determinare il tipo di distanza e il numero di vicini più adeguati. I risultati ottenuti sono incoraggianti e in linea con quelli menzionati in una vasta letteratura ed in innumerevoli studi, soprattutto prodotti da Paesi nordici. Quando le stime è stata elaborata a livello di ciascun pixel, ne sono derivati errori relativi elevati, dal 76 al 99%. Quando i valori medi dei pixel dentro de l'unità de campionamento con 30 metri di "buffer" è stato utilizzato per derivare le variabili forestali, l'errore relativo si è abbassato a 47-63%. Gli errori ottenuti sono piuttosto elevati rispetto a quelli ammessi dagli inventari forestali (di solito inferiori al 20%) e non sono sufficientemente precisi per la realizzazione di piani di gestione forestale del distretto e delle concessioni forestali. Comunque, la spazializzazione di attributi forestali tramite *k*-NN con l'errore noto è un potente strumento per le determinazione di decisioni strategiche. Le stime delle variabili forestali ottenute con questo metodo possono essere facilmente prodotte e usate in aree dove nessuna altra informazione sia presente o per aree prive di campioni.

Parole chiave: telerilevamento, k-NN, inventario forestale multifonti, foreste miombo

List of acronyms

BA - Basal area (m^2 ha^{-1})
Dbh - Diameter at breast height (cm)
DN - Digital number
ETM - Enhanced thematic mapper
FIU - Forest inventory unit
Hc - Commercial height
Ht - Total height
Mir - Middle infrared band
MSS - Multi spectral sensors
NDVI - Normalized difference vegetation indice
Nir - Near infrared band
RMSE - Root mean standard error
SD - Standard deviation
TM - Thematic mapper
Vt - Total volume (m^3 ha^{-1})
Vc - Commercial volume (m^3 ha^{-1})

List of figures

List of tables

Definitions

Albedo: is the fraction of solar energy (shortwave radiation) reflected from the Earth back into space.

Bit: is the short term for binary digit, which can take two possible binary values, 0 or 1. It is the smallest unit of storage and information within the computer.

Digital image: a digitally encoded record of spectral reflectance or emittance intensity for a selected object or area. A digital image is usually composed of one or more spectral bands. For each band, an individual pixel will have a digital number value (DN) to represent the intensity of spectral reflectance of the objects represented by the pixel.

Digital number: a positive integer representing the relative reflectance or emittance of an object in a digital image. For a 8 bit image the DN value lies in the range 0-255.

Electromagnetic energy: is a dynamic form of energy (visible light, radio waves, heat, X- rays, ultra-violet rays, microwaves,…) that is caused by the oscillation or acceleration of an electric charge. All natural and synthetic substances above zero degrees continuously produce and emit electromagnetic energy in proportion to their temperature. Electromagnetic energy passes through space at the speed of light in the form of sinusoidal waves. The wavelength is the distance from wavecrest to wavecrest.

Electromagnetic spectrum: is a continuum of all electromagnetic waves arranged according to frequency and wavelength.

Forest types: a category of forest stands defined by composition, structure and age for management purposes. Saket (1995) defined the following forest types for Mozambique base on structural characteristics:

> LF1 – lowland forests dense forests : vegetation with crown cover of upper stratum over 70 % with trees with more than 7 meters height and a herbaceous layer poorly developed.
>
> LF2 – medium closed lowland forests: Vegetation composed of two to three strata. The crown cover of the dominant strata ranges from 40 to 70 percent. Trees are taller than 7 meters height and a moderate dense shrub layer is present with a sparse ground stratum. Herbaceous strata may or may not be present.
>
> LF3 – Open lowland forests: vegetation with upper stratum crown coverage between 10 to 40 percent. The vegetation is composed of one to two tree strata associated with a moderate to abundant grass layer and a medium to dense shrub layer. A herbaceous stratum is present.
>
> T- thicket : vegetation with predominance of a tree layer and shrubs with crown cover percentage less than 10%. Is frequently the result of a degradation process following burning, overexploitation and overgrazing of forests.
>
> S-shrubland: it is characterized by a low shrubby stratum (50cm to 3 meters) with occasional emergent trees with heights superior than 7 meters. Crown coverage percentage is variable (from 10% to high).
>
> WG – wooded grassland: vegetation characterized by the predominance of a grass layer and a tree coverage less than 10% with variable tree height.
>
> G – grassland: vegetation where predominates all types of herbaceous layer with some scattered trees. It is very difficult to distinguish and separate from agriculture.

A- agriculture: area where agriculture are being carried out within any of the other forest types, including grass covered lands previously used for agriculture and left as fallow.

Multisource forest inventory: uses various sources of geo-referenced data in addition to field inventory plots.

Pixel (combination of Picture & Element): is the smallest element of a display which can be assigned a colour.

Radiation: is the transfer of energy from one body to another in absence of an intervening medium. It is the only method by which solar energy cross space and enter the earth's atmosphere.

1. Introduction

1.1 Remote sensing fundamentals

Understanding the basics of remote sensing is the prior step to understand the methodology used in this study, and the fundamentals of remote sensing science are generally described in this thesis chapter.

Remote sensing is initially defined as the collection of data about objects which are not in contact with the collecting device (Howard, 1991). The historical development of remote sensing has been divided into two general phases: (1) prior to 1960 when predominates the use of aerial photography and (2) after 1960 when started the use of satellites as the space platforms (Howard, 1991).

The evolution of remote sensing and its applications in forestry is connected not only to the evolution of data collection devices and platforms (balloons, aircrafts, satellite sensors and space shuttles) as well as to data processing, imagery analysis and presentation technology development. Consequently, remote sensing definition has evolved from "reconnaissance from a distance concept" to a more broad definition of reception, pre-processing, interpreting and latter analysis of data obtained from orbiting satellites (Campbell, 2002; Franklin, 2001).

There are several satellites types orbiting around the earth that provide a myriad of earth images. This study uses Landsat 7 ETM+ images and a general description of this satellite is given in this chapter. In general, satellites can be categorized according to its orbit into two main groups: (a) *geostationary satellites*, that circle the Earth in a geo-synchronous orbit over the equator, that is, they orbit the equatorial plane of the Earth at a speed matching the Earth's rotation allowing the Earth observation from the exact same place all the time and therefore continuously monitor a single position on the earth's surface. They give very coarse data (low spatial resolution > 2500 m/pixel) due to the fact that they usually orbit very far from earth (about 36 000 km above earth). There are more than 300 geostationary satellites mostly for meteorological and communications purposes, such as Meteosat (Europe), GOES (America), GMS (Japan/Australia), IODC (Indian Ocean), Skynet, etc. (b) *polar orbiting satellites*, constantly circling the Earth in an almost north-south orbit, passing close to both poles, sun synchronized, that is, covering each area of the world at a constant local time of the day known as local sun time. Their orbits are near-polar, covering the latitudes 82° north to 82° south of the equator with orbiting altitudes varying between 700 to 1500 km. Landsat and SPOT series of satellites have been the best-known polar-orbiters and have provided the most commonly used data in Mozambique.

The underlying basis for most remote sensing methods is the measurement of the *electromagnetic energy* emanating from distant objects made of various materials, so that we can identify and categorize these objects by class or type, substance and spatial distribution. The electromagnetic energy is then captured by sensors which are classified in two categories: (a) *active sensors* when they transmit electromagnetic radiation and measure the reflection of this radiation such as the radar systems, and (b) *passive sensors* that sense naturally available energy such as the sun and do not output any form of electromagnetic radiation

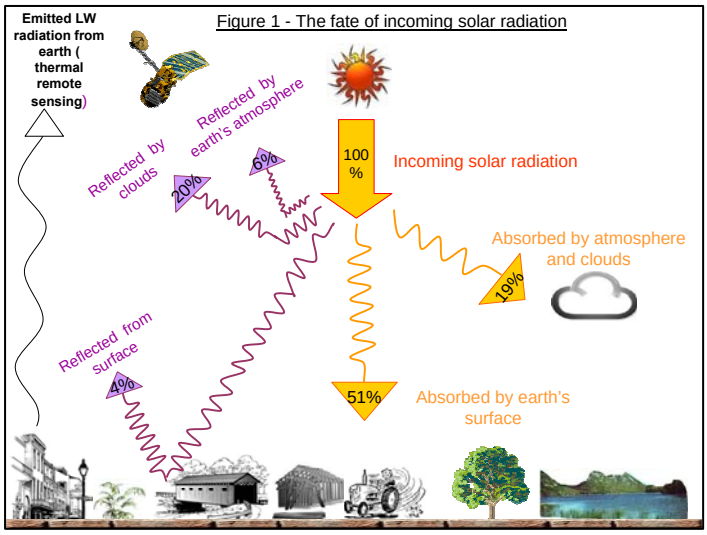


Figure 1 - The fate of incoming solar radiation

towards the target area. The sun and the earth's ambient temperature, (such as soil, water and vegetation) also known as the "terrestrial thermal infrared energy ", are the most common sources of radiant energy used in remote sensing (Lillesand and Kiefer, 1979).

Of the total income solar energy, 51% is absorbed at the surface, 19% absorbed by atmosphere and clouds and only 30% is reflected by the atmosphere, clouds and surface, that is, the albedo (Figure1).

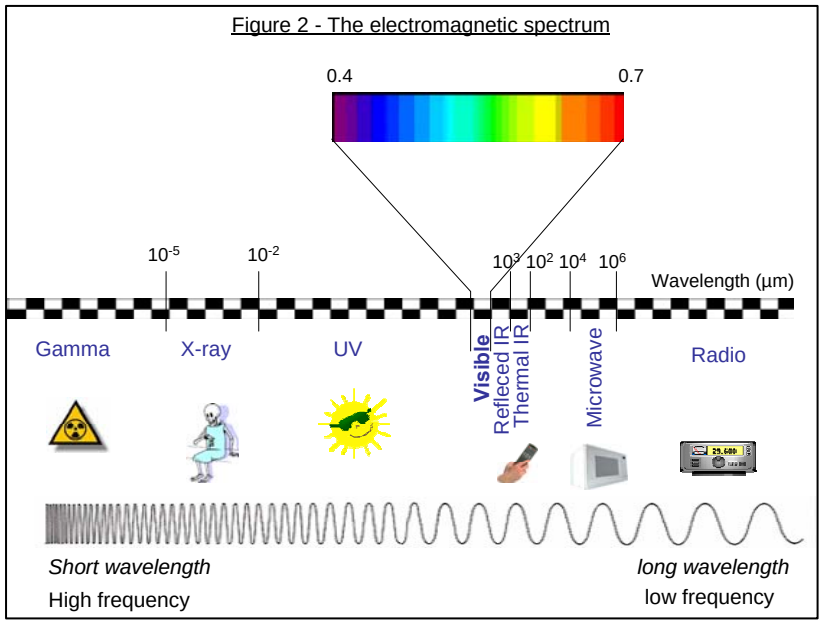
Therefore, the electromagnetic radiation captured by sensors is affected by two main mechanisms: (1) the atmospheric scattering /diffusion and absorption mechanisms and (2) the reflective characteristics of ground surface. When the radiant energy passes through the atmosphere, the amount of energy available to be detected by the sensor is affected by the atmospheric scattering, that is, the unpredictable diffusion of radiation by the existing particles in the atmosphere (gases, pollen, dust, smoke, water droplets…) that change the spatial distribution of the energy. On the other side, the atmospheric absorption is the process by which incident radiant energy is retained by the atmosphere constituents (water vapour, carbon dioxide and ozone). These gases absorb a portion of the radiant energy that is then re-emitted, usually at longer wavelengths, and some of it remains and heats the atmosphere;

When electromagnetic radiance from the sun, that is not absorbed or scattered in the atmosphere, hits the earth's surface three fundamental energy interactions take place: reflection, absorption and transmission. *Reflection* reflects light what we know as colour; i.e. chlorophyll in plants reflects green light; *absorption* takes place when the incident energy that is not reflected or transmitted is transformed into another form, such as heat or absorbed by chlorophyll in the process of photosynthesis; *transmission* takes place when energy propagates through a medium. The principle of conservation of energy points out that the incident energy is equal to the sum of reflected, absorbed and/or transmitted energy

On the other hand, when the electromagnetic energy hits the ground the reflective characteristics of the ground are different for each earth feature allowing us to distinguish different features on an image. Different materials reflect light of different wavelengths and by different amounts, known as *spectral reflectance curve or spectral signature*. When viewing a remote sensing image on a particular band we see a greyscale image showing the electromagnetic radiance received in that band from each point on the ground. Areas of higher radiance show lighter shades and areas of lower radiance show darker shades.

One advantage of satellite images when compared to aerial photographs is the possibility to capture non visible portions of the *electromagnetic spectrum* such as the Near Infra-Red wavelength. Wavelengths are measured in micron and the spectrum of waves is divided into sections based on wavelength. The *shortest waves* are gamma rays, which have wavelengths of 10^{-6} microns or less. The *longest waves* are radio waves, which have wavelengths of many kilometres. The light used to "see" an object must have a wavelength about the same size as or smaller than the object (Figure 2).

Figure 2 - The electromagnetic spectrum

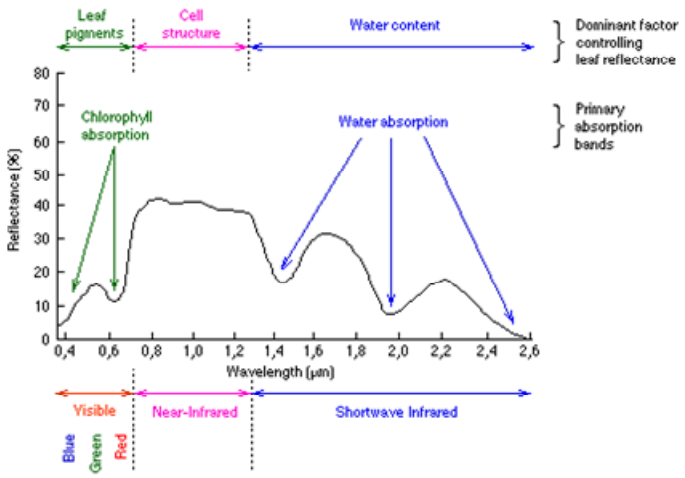


1.1.1 Spectral properties of vegetation

This property is of most interest for choosing the appropriate wavelength region for a particular application. The spectral reflectance curves for healthy vegetation manifest "peak and valley" configuration (Lillesand and Kiefer, 1979). Vegetation reflects much more energy in the near-infrared (0.8 to 1.4 microns) wavelength than it does in visible light (0.4 to 0.7 microns) as it is simply portrayed in figure 3.

The spectrum of plant leaves can be divided into three distinctive ranges:

Figure 3 - Typical spectral response characteristics of green vegetation.



Source:Crum, Shannon; http://www.rsunt.geo.ucsb.edu/rscc/vol2/lec2/2_2.html #21

1 - the range between 0.4 µm and 0.7 µm (visible band) is characterised by very low reflectance due to intense absorption of the incident radiation by pigments in the plant. On the visible portion of the solar spectrum, chlorophyll controls much of the spectral response of the living leafs (Campbell, 2002). Chlorophyll pigments absorb at 0.43 - 0.45 µm (blue), and have an additional absorption band at about 0.65 µm (red) . There is a slight increase in reflectivity around 0.55 µm band (visible green) because the pigments are least absorptive there.

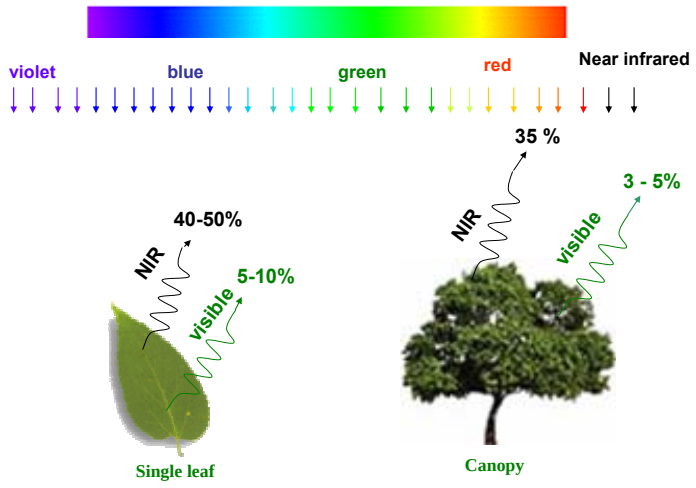
Only 5-10% of the incident energy of the visible portion is reflected by a single leaf. This reflection makes leaves to appear green specially in the summer when the content of chlorophyll is at its maximum. During autumn/winter there is less chlorophyll content in the leaves and therefore they absorb less green wavelengths and reflect more red light making them to appear reddish or yellow.

2 – In the near infrared portion of solar spectrum (between 0.7 µm and 1.3 µm) there is very little absorption and high reflectance. The reflectance of the leaf is controlled by the structure of the spongy mesophyll tissue, which causes multiple reflection of near infrared radiation at cell walls.

3 - The range between 1.3 µm and 2.6 µm (short infrared wavelengths) is characterised by pronounced minima. Leaf water content appears to control the spectral properties of the leaf, that absorbs most energy at these wavelengths as it is illustrated in figure 3. Reflectance of leaves is approximately inversely related to leaf liquid water content. This property can be used for identifying tree types and plant conditions (water stress) using remote sensing images

Vegetation canopies are composed of many separated leaves with several sizes, shapes, orientation and coverage of ground surface (Campbell, 2002). In a forest canopy upper leaves form shadows that cover the lower leaves and the overall reflectance is a combination of leaf reflectance and shadow. Campbell (2002) cited a study carried out by Knipling (1970), that illustrates that the canopy reflectance is less than the values observed in the laboratory for individual leaves. The reduction in reflectance for canopies is although much lower on the near infrared region than on visible spectra (figure 4).

Figure 4 - Reflexive properties of single leaf and canopies – Percent reflected

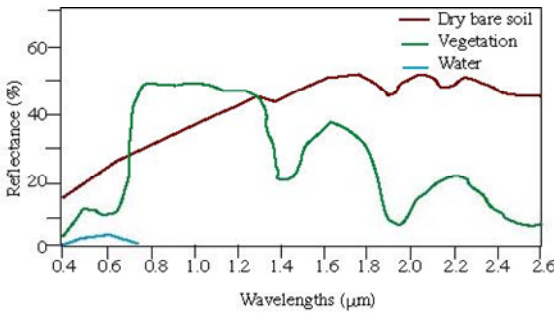


Source: Campbell, 1996

The reflection of incident energy depends also on the vegetation cover. If vegetation cover is less than 100%, as it is in most of the cases for African savannas forests, then the reflectance properties of the underlying soil and water will also be also recorded by the remote sensor.

The difference of reflection in red band and NIR band is great for vegetation areas but insignificant for bare soil. Another important feature is that clear water reflects visible light only, so it will appear dark in infrared images (figure 5).

Figure 5- Typical spectral reflectances curves for vegetation, soil and clean water



Source:Crum, Shannon; http://www.rsunt.geo.ucsb.edu/rscc/vol2/lec2/2_2.html#21

Due to different reflectance curves the band ratioing (which underlies the vegetation indices construction) in the visible and infrared spectra can be a useful tool for imagery analysis.

Spectral vegetation indices

Spectral vegetation indices expresses the remotely sensed response observed in two or more wavebands as a single value that is related to the biophysical variable of interest (Foody *et al*., 2001 quoting Mather, 1999). They have been used since the late 1960's and are continuously evolving and more than 50 vegetation indices have been developed for different applications and to account for noises in reflectance (Bannary *et al*., 1995). Healthy vegetation absorbs visible light, mostly red light, and reflects near-infrared. If a pixel has a large difference between NIR and Red band we assume that

the pixel probably represents an area covered by vegetation. A second assumption is that all bare soils in an image will form a line in spectral space (a red-near-infrared line for bare soils). This line is considered to be the line of zero vegetation.

Vegetation indices are classified accordingly to the orientation of lines of equal vegetation (isovegetation lines) in relation to the bare soils line: (a) *slope indices*, which consider that the isovegetation lines converge at a single point and the vegetation indexes measure the slope between the point of convergence and the red-NIR point of the pixel (e.g: NDVI, RVI); (b) *perpendicular distance indices*, which consider that all isovegetation lines remain parallel to the soil line and measure the perpendicular distance from the soil line to the red-NIR point of the pixel (eg. PVI).

Some most used indices are below described. They can be classified as:

(1) Simple ratio vegetation index, such as the ratio vegetation index (RVI) attributed to Jordan (1969) ranges from zero to infinite and indicates the amount of vegetation by:

$$RVI = NIR\ band / Red\ band \qquad \text{[eq.1]}$$

(2) Normalized difference vegetation index, being the most widely used the Normalized Difference Vegetation Indice described by Rouse et *al.* (1973) though the concept was discussed by Kriegler *et a*l. (1969) as :

$$NDVI = \frac{NIR - RED}{NIR + RED} \qquad \text{[eq.2]}$$

NDVI values vary from -1 to +1 . When there is no vegetation (zero percent cover) NDVI varies from 0 (snow) to 0.2 (organic black soil). If the NDVI it is between +0.2 to +0.3 correspond to shrubland/grassland area or open forests that are considered intermediate green cover. When green cover is between 30-40% the soil noise becomes at a maximum and decreases up to zero when a 100% green cover is achieved (NDVI between 0.6-0.8 indicates an area of tropical forests). A negative NDVI indicates an area of clouds, water or snow. Despite of the major weaknesses of NDVI (sensitive to variations in the background signal due to soil color, moisture and litter and influenced by atmospheric scattering) the widespread and strength of NDVI lies on its relative stability due to its power to reduce many form of noises such as: cloud shadows, topographic variations, illumination differences.

(3) Distance-based vegetation indices are designed to eliminate the effect of the background soil brightness and only detect the features of the vegetation cover based on Perpendicular Vegetation Indice (PVI) soil line concept which is a linear regression of the near infrared band against the red band for pixels of bare soil. PVI is the Perpendicular Vegetation Indice which was first described by Richardson and Wiegand (1977) as follow:

$$PVI = \sin \alpha\ (NIR) - \cos \alpha\ (red), \qquad \text{[eq.3]}$$

where α is the angle between the soil line and the NIR axis.

The vegetation indexes above described assume that there is soil line and a single slope in red-NIR space. But not all soils are alike and soils have different red-NIR slopes in a single image as well as changes in soils moisture cause noise to the vegetation reflectance. This problem is mostly detected when vegetation cover is low. Hence, other vegetation indexes were refined to account for the spectral reflectance properties of the underlying soil. The Soil Adjusted Vegetation Index (SAVI), designed by Huete (1988), acknowledges that the isovegetation lines are not parallel and that they do not all converge to a single point and that there is a simple radiative transfer from soil to vegetation. The index is based on a correction factor that varies from 0 for very high vegetation densities to 1 for

very low vegetation densities areas, and the standard value commonly used for intermediate vegetation density areas is 0.5

$$SAVI = \frac{NIR - red}{NIR + red + L}(1 + L), \qquad \text{[eq.4]}$$

where *L* is a correction factor which ranges from 0 for very high vegetation densities to 1 for low vegetation densities.

This indice is slightly less sensitive to changes in vegetation cover than NDVI (but more sensitive than PVI) at low levels of vegetation cover.

(4) Indices that account for reducing atmospheric noise are also being developed. Kaufman and Tanre (1992) introduced the atmospheric corrections in the Vegetation indice called Atmospherically Resistant Vegetation Indice (ARVI). The red reflectance variable in NDVI was replaced with the term:

rb = red - gamma (blue - red), where the value of gamma is 1.0

$ARVI = \frac{NIR - rb}{NIR + rb}$, and ranges from -1 to +1. [eq.5]

Spectral relationships and forest inventory variables

Franklin (2001) mention that there has been several studies done to relate forest parameters to satellite data , and the relationships found are similar to those of aerial photographs:

- Increasing crown development is reflected by decreasing visible reflectance (darker tones);
- Larger crowns absorb more light and reflect more the infrared;

The same author refers that remote sensors can only detect data that directly influence on the spectral response: illumination geometry (that is, target - sensor solar conditions) and amount of vegetation viewed from above (crown cover and canopy top reflection). In high resolution images (of 1 meter order) sensors detect several measurements for each tree, being each pixel records dependent on the part of the tree crown from which it was acquired. In Thematic Mapper Landsat images each ground resolution cell (nearly the size of a pixel) represent the average spectral response from several trees (Wallerman, 2003).

Remote sensing has proved useful for extracting land cover/use classes (forests versus non forests) and vegetation types/ecosystems defined by attributes detectable by remote sensors (for example based on crown cover percentage) and therefore for stratification and forest inventory planning (Franklin, 2001). Forest inventory variables such as dbh, commercial and total height, volume and so on, are considered second or third order variables, that is, can only be indirectly estimated through the existing correlations between crown size and closure and the indirect variable (Franklin, 2001).

Franklin (2001) cited a study carried out by Ripple *et al*. (1991) where the spectral response of Landsat TM data in homogeneous Douglas fir stands in USA was analyzed. It was found an inverse curvilinear relationship between the volume (m^3 ha^{-1}) and many of the spectral bands (digital number), explained by the fact that young coniferous stands have a high reflectance due to a highly reflective presence of shrubs and herb layer while older stands with higher volume and shadows presented low reflectance.

On more heterogeneous and fragmented landscapes in Greek Mediterranean occurring patches of forested land mixed with agricultural farms and woodlands, stand volume and basal area show limited correlation with image data when compared to similar studies in uniform boreal forests due to many factors affecting the spectral response of vegetation (Mallinis *et al*., 2004). The following main factors affect the TM data ability to predict volume (Franklin, 2001):

- The homogeneity of forest stand.
- The spatial scale.
- The range of volume to be investigated.

Although many studies were made in the past decade to identify the relationship between stand parameters and spectral response, the conclusions are site specific and the influence of different study areas characteristics on the spectral response is poorly understood and therefore conclusions can only be indicative. The following general conclusions can be extracted from the different studies (Lu *et al,*. 2004):

- visible bands are strongly related to biomass;
- middle infrared band has a strong negative relationship with stand timber volume;
- middle infrared bands are most sensitive to changes in forest volumes with reflectance in this spectral region being strongly related to canopy closure (Horler and Ahern, 1986);
- shadowing controls largely the forest canopy spectral response (Ardö, 1992);
- The near infrared wavelength reflectance present all kind of relationship with stand parameters: positive, negative or flat , due to shadowing and decrease of soil brightness;

In boreal forests of Canada, Horler and Ahern (1986) concluded that the Landsat TM band 3 (red), TM band 4 (NIR) and TM band 5 (mid infrared) were the most appropriate bands for forest-cover-type discrimination.

Lu *et al.*(2004) studied the relationship between forest stand parameters and Landsat TM spectral responses in three sites of Amazonian basin and concluded that middle infrared band (TM5) was strongly correlated with forest stand parameters (biomass, basal area, mean diameter and height) and independent of biophysical environment.

1.1.2 Main characteristics of satellite sensors

Geostationary and polar orbiting satellites provide a large selection of remotely sensed data with respect to four parameters: (1) spatial, (2) temporal, (3) spectral and (4) radiometric sampling of earth surface. Its application and suitability depends on the resolution and extent of all 4 data parameters given by the satellite sensors. *Resolution* refers to the intensity or rate of sampling and is related to the smallest feature while *extent* refers to the overall coverage of a data set and is expressed as the largest feature.

Satellite remote sensing data captured by the sensors can be characterized by four main parameters:

(1) *Spatial resolution* which refers to each individual pixel size and therefore expresses the level of detail depicted in an image. In practice, represents a measure of the smallness of objects on the ground that may be detected as separate entities in the image. Pixel size is a function of the sensor characteristics (optics and sampling rate) and of platform (altitude and velocity) and vary in different satellite types and bands. The greater the distance between the satellite and earth the lower spatial resolution of image is obtained (for example, Quickbird = 450 km; Landsat = 705 km , Spot = 832 Km). *Spatial extent* refers to the overall image coverage also referred as swath width across the satellite orbit.

(2) *Temporal resolution* which indicates how often a sensor is acquiring an image for the same particular area. This imaging on a continuous basis at different times provide the possibility to monitor changes that take place on earth's surface and are specially important if persistent clouds refrain the clear view of earth's surface, short-lived phenomena are needed to be monitored and multi-temporal comparisons are required. Temporal resolution of a sensor depends on many factors, such as, sensors capabilities, swath overlap and altitude and is expressed as the number of days needed to repeat the orbit cycle. Temporal resolution varies from 2 images per day (NOAA), 1 image every 3 days (Quickbird) up to 1 image every 26 days (SPOT). Higher the temporal resolution there better chances of cloud free coverage. *Temporal extent* refers to the total period over which imagery is available, since satellites have an average life period from 3-5 years and can suffer some equipment disruption.

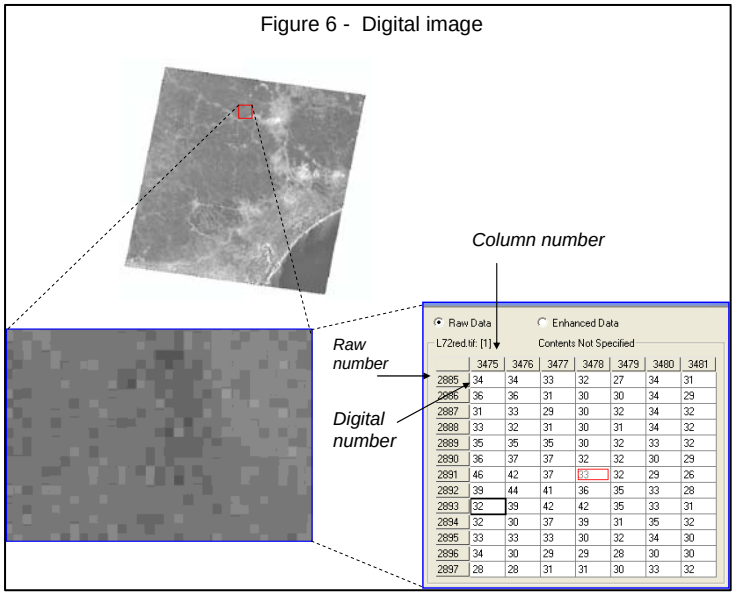
(3) Spectral resolution refers to the ability of a sensor to define fine wavelength intervals and is expressed as the wavelength width. Spectral resolution is an important image feature since it contributes to distinguish different targets based on their spectral response in each band. Panchromatic photographs that capture one wide band (0.4 - 0.7 µm) are considered having low (coarse) spectral resolution, while Landsat is considered to have high spectral resolution since its sensors are sensitive to the individually reflected energy as blue, green and red wavelengths. More recently, advanced multispectral sensors also known as hyperspectral detect hundreds of very narrow spectral bands through the visible, near-infrared and mid-infrared portions of the electromagnetic spectrum (for example MODIS is considered to have a hyperspectral resolution (21 bands within the UV to NIR, 15 bands within the TIR, all with narrow bandwidths). *Spectral extent* refers to the number and spectral range of channels in the image.

(4) Radiometric resolution refers to the ability of the sensors to discriminate very slight differences in emitted or reflected energy and is expressed in terms of bits. Image data are displayed in a range of grey tones, with black representing a digital number of 0 and white representing the maximum or the saturation value (for example, 255 in 8-bit image). Higher radiometric resolution means more shades of grey, expressed by the number of the possible range of values a pixel may obtain. Landsat MSS detectors pixel range up to 128 levels (7-bit image), while Landsat TM pixel ranges up to 256 (8 bit image). *Radiometric extent* expresses the range of brightness values a sensor band is set up to capture and is expressed as the maximum brightness.

1.1.3 Characteristics of digital image

While aerial photography produces analogue images of the terrain, that is, an image represented by continuous variation in tone, images obtained by satellite remote sensing are digital. Digital images are reduced into numbers arranged in two-dimensional matrix (or raster data) of individual picture elements, known as pixels (figure 6). When a image is represented as numbers, brightness can be manipulated. Digital images have advantages over analogue film images because computers can store, process, enhance, analyze, and render images visible on a computer screen (Campbell, 2002).

Figure 6 - Digital image



Each pixel represents an area on the earth's surface and stores two information: (1) the intensity value or pixel value and (2) location address numbers (given by its row and column number).
The *address* of a pixel is denoted by its row and column coordinates in the two-dimensional image. There is a one-to-one correspondence between the column-row address of a pixel and the geographical coordinates (e.g. Longitude, latitude) of the imaged location.

The *intensity value*, also known as the *pixel value* or *grey level,* represents the measured physical quantity such as the solar radiance in a given wavelength band reflected or/and emitted from earth's surface. This value is normally the average value for the whole ground area covered by the pixel and is affected by the following main factors (Franklin, 2001):

- the spectral properties (reflectance, absorption, transmittance) of the earth surface targeted;
- the illumination geometry and topographic effects;
- the atmosphere;
- the radiometric properties of the sensor, and
- the geometrical properties of the target (e.g., leaf inclination).

The intensity value or pixel value is digitised and recorded as a digital number. Due to the finite storage capacity, a digital number is stored with a finite number of bits (binary digits) which express the radiometric resolution of the image. The detected intensity value needs to be scaled and quantized to fit within this range of value. In a radiometrically calibrated image, the actual intensity value can be derived from the pixel digital number.

The use of digital number (DN) facilitates the visual display of image data and the design of instruments and data communication (Campbell, 2002). When doing analysis that compare relative brightness the use of digital numbers is satisfactory while for comparisons from scene to scene the use of radiances is more adequate since DN from one scene does not represent the same brightness as the same DN from another scene. Literature review shows that when using *k*-NN method to expand forest inventory variables from plot level to Landsat scenes, digital numbers directly extracted from the images data have been mostly used.

Digital remote sensing data is usually organized and stored into three main formats (Campbell, 2002) : (1) band interleaved by pixel (BIP) where data is organized in sequence values for each band by pixel; (2) band interleaved by line (BIL) where data is organized for each line on each band and (3) band sequential (BSQ) format where data is organized in sequence for each band treated as a separated unit.

1.1.4 Fundamentals of image processing

Once the raw remote sensing digital data has been acquired, it must be processed into usable information form. Computer digital image processing procedures can be categorized under the following main operations (Lillesand and Kiefer, 1979):

- ***image restoration*** that correct image distortions in order to truly represent the targeted object and involve mostly radiometric and geometric distortions corrections,
- ***image enhancement***, that is the enhancement of image features through several techniques such as contrast and stretching in order to prepare the data to be easily interpreted, and
- ***image classification***, which deals with quantitative techniques that automatically interpret the digital data and assign to each pixel an information category.

Image restoration, which comprehend radiometric and geometric corrections are generally done by the image provider before delivering the data to the costumer and are also known as the pre-processing operations.

Landsat 7 ETM has onboard sensor radiometric correction for displacement of the scanner sweep with the forward movement of the satellite of ± 5%. Image based radiometric corrections are made to the

raw digital image data to remove for atmospheric effects from the image. The radiance values registered by the sensors do not directly reflect the ground reflectance because of the light scattering of a constantly changing atmosphere, also known as haze. Haze has an additive effect resulting in higher digital number values, decreasing the general contrast of the image. Haze is also wavelength dependent and is mostly pronounced in shorter wavelengths and negligible in near Infrared band.

There are three main methods to estimate the atmospheric scattering: (a) image-based method which comprises the histogram minimum method and regression methods, (2) radiative transfer models and (c) empirical line model. Image-based methods are simple and do not require data about atmospheric conditions at the time image was taken like the radiative transfer models or field radiometers to measure the ground reflectance of a dark and bright target like in the empirical line method. The simplest atmospheric correction is performed by relating image information to pseudo-invariant reflectors, such as lakes, dark asphalt and rooftops (Franklin, 1991). Lillesand and Kieffer (1979) describe the histogram minimum method based on the assumption that pixels with the lowest digital number in the image dataset represent the features with nearly no reflectance (e.g., dark coloured rocks, water bodies) and therefore representing only the atmospheric scattering. If the average value of the lowest digital number pixels is subtracted in all dataset, the atmospheric scattering is eliminated and this operation is also called haze removal. On the other hand, the regression method is similar and applicable to dark pixels. The NIR pixel values are plotted against pixel values in the other bands. Regression methods best fit the line to data plotted and the resulting bias is assumed to represent the atmospheric radiance, and therefore subtracted from the all image.

The topographic effect is defined as the variation in radiance from inclined surfaces, compared with radiance from a horizontal surface, due to the surface orientation in relation to the light source and sensor (Franklin, 1991). The same author states that topographic corrections are more difficult to correct than haze removal and results are comparatively less certain.

Radiometric correction is specially needed if multitemporal images analysis are required as well as when more than one image is needed to cover a study area, or data is provided by different platforms (Howard, 1991). By removing haze from the image, the brightness values across the set of images are standartized making subsequent analysis easier and more reliable .

Geometric corrections are undertake to reduce the inaccuracy between the location coordinates of the targeted elements in the image data and the actual location coordinates on earth ground, so that the imagery will be as close as possible to the real world. This corrections are due to changes in altitude (topography), velocity of the satellite, eastward earth's rotation and earth's curvature (Howard, 1991 citing Berstein and Ferneyhough, 1975). Geometric correction involves the *resampling* of the image.

This process consists of two main stages:

(1) The first stage concerns the selection of a set of ground control points (with known UTM coordinates) that are easily identified on the image. Usually the UTM coordinates are obtained from a topographic map in either digital (Digital Earth Models) or paper form and the image points must be matched to the ground control points. When considering the selection of ground control points it is important to consider two aspects: (a) the number of control points and (b) location of control points, which should be dispersed throughout the image with good coverage near the edges. Regarding the number of control points, Campbell (2002) quoting a study by Bernstein *et al*. (1983) recommended the use of 16 control points as a reasonable number since the accuracy of the ground control points may decrease as the number of control points increase because the analyst usually chooses the best points first. 16 ground control points can be considered acceptable if each of them can be located with an accuracy of one third of a pixel, considering as well that they are well distributed and the nature of landscape on the study area.

Two sets of data is obtained for each ground control point: their image coordinates and their true ground coordinates. These two sets of data are used to establish a mathematical relation through regression that interrelate the geographic and the image coordinates.

(2) The second stage consists on deciding the mathematical algorithm that will be used to assign the digital number to the newly geographic oriented pixel (usually in UTM coordinates). Three sampling algorithms are commonly used: nearest neighbour where the digital value for the corrected image pixel is derived from the digital number of the nearest pixel of the original image. This method is simple and does not change the original pixel values. The bilinear interpolation resampling algorithm uses the proximity-weighted for the 4 nearest pixels in the original image around to the output pixel location. This may result in changes to the original pixel value creating a new digital value in the geometric corrected image. The cubic convolution resampling algorithm which weighted averages the 16 closest pixels from the targeted pixel. This method results as well, in completely new pixel values but produces the sharpest image (Lillesand and Kiefer, 1979).

Image enhancement refers to procedures that simply transform the data into a more interpretable form and are classified as point and local operations (Lillesand and Kiefer, 1979). Point operations refer to procedures that modify the brightness values of each pixel in an image data set independently while local operations modify the value of each pixel considering the context of the brightness values around it, or neighbourhood.

Image classification is the process of assigning pixels to classes, that in theory are homogeneous, that is, those pixels within classes are spectrally more similar to one another than they are to pixels in other classes. Image classification is an important component of pattern recognition and image analysis field. There are many classification strategies that have been developed over the years but the most common ones can be referred as (Campbell, 2002): (a) point classifiers where each pixel is individually assigned to a class according to the several values measured in separate bands and (b) spatial or neighbourhood classifiers where groups of pixels are classified using both spectral and textural information. Classification has also been referred into two distinct types, according to the interaction with the image interpreter and field data (Campbell, 2002): (a) supervised classification which field data with known classes provide a guidance to the classification and (b) unsupervised classification where the image analyst has minimal interaction and classification is automated based on spectral classes that are uniform in respect to brightness.

When combining forest inventory sample data to pixels in the image to extrapolate the information at pixel level to the whole image, such as the one that is performed in this study, a supervised classification is used. The classifier is a computer programme or algorithm that implements a specific procedure for image classification. Supervised classification has been simply defined as the process of using samples of known identity (also known as training data) to classify pixels of unknown identity (Campbell, 2002). The main advantages of supervised classification are summarized as:

a) supervised classification allows the analyst to control the informational classes tailored to specific purposes or region. Unpredictable classes are not generated.
b) Supervised classification is tied to the geographic area from where the training data was collected.
c) Image analyst can detect errors in classification by using a validation dataset to determine the inaccuracy of classification.

Since supervised classification is based on training data which is collected with the aim of identifying a set of pixels that accurately represent spectral variation present in each class, the key characteristics of training areas are summarized as follows (Campbell, 2002) :

- *Number of pixels*: training areas should provide at least a total of 100 pixels for each class.
- *Size:* small training areas are difficult to locate accurately on the image while too big training areas tend to include undesirable variation. The author recommend each analyst to define the

optimum size of individual training samples since it varies according to the heterogeneity of each landscape and class.

- *Shape*: shape of training areas in not considered important.
- *Location:* training data must represent variation within the image and therefore must not be clustered in favoured classes or regions.
- *Number:* the optimum number depends upon several factors. A selection of a large number of training samples is recommended to allow for discard of some of them, as well as, the common sense indicates that is usually better to have many small training areas than use only a few large areas.
- *Placement:* training areas should be placed so that accurate location on image is allowed and should not encompass edge pixels of the class.
- *Uniformity*: it is considered the most important characteristic and data in each training area should exhibit a unimodal frequency distribution for each spectral band to be used.

1.2 Application of remote sensing in forest inventories and management

Forest inventories and field data collection can be supported either by airborne imagery (aerial photographs) or by spaceborn imagery (satellite images). Many authors review with several scales of details the application of aerial photographs for forestry science with emphasis on forest mensuration, timber cruising, forest management planning and monitoring, timber harvesting, roads planning, forest health and damages assessments (Loetsch and Haller, 1973; Lillesand and Kiefer, 1979). The use of aerial photography in forest inventories goes back to 1920's when it was firstly used in Quebec boreal forests stock map preparation which combined aerial photo-interpretation and field checking plots (Howard, 1991, citing Wilson 1920). Today, the application of aerial photographs have been largely demonstrated either for both strategic inventories of large areas and for management purposes inventories of stand areas (Loetsch and Haller, 1973; Schreuder *et al*., 1993; Tuominen and Poso, 2001).

On the other hand, the launch of US Earth Resources Technological Satellite (ERTS-1) in 1972, latter renamed Landsat 1, marks the effective application of satellite remote sensing in forestry (Howard, 1991). Landsat satellites launched with a sun-synchronised and repetitive orbit (every 18 days) at nominal altitude of 900 km provide a fast and continuous information of large earth areas (around 3 million hectares per image) offering vast applications in other sciences than meteorological forecast.

Literature cites several advantages of using remote sensing in forestry, particularly in large area surveys and for monitoring broad scale changes (continental, regional, national) in forest cover, density and land use (Loetsch and Haller, 1973; Howard, 1991). Boyd *et al.* (2002) citing DeFries and Townsend (1999) refer that remote sensing provides geographical referenced data for land cover mapping more accurate than data collected only by field surveys.

Multisource forest inventories employ various sources of geo-referenced data, in addition to field inventory plots in order to obtain more reliable estimates or estimates of smaller areas than when employing the pure field plot data only (Katila, 2004).

Besides the advantage of obtaining georeferenced data, the immense image archives from past decades of remotely sensed data acquired by earth resources satellite systems provide an unprecedented dataset to study land cover transformations over time and space (Boyd *et al*., 2002). When this archives are complemented with the increased resolution from satellite platforms and the reduction of image acquisition costs the use of remote sensing data for land management and monitoring purposes at various spatial scales is enhanced (Chen and Rogan, 2004).

Forest inventories use forest stratification into smaller, relative homogeneous forest types/stands in order to reduce variance and improve the precision of sample estimations. In statistical terms, forest

stratification should have minimized within-stratum variance while maximizing between-strata variance. In ideal terms, a forest type will be perfectly homogeneous and extremely different from the other defined forest types. In natural forests is not often the case since there are no clear borders between different forest types.

Forest stratification is often done by aerial photographs and more recently with the availability of satellite images, it has been done with the help of vegetation indexes and supervised image classification. Literature mention several advantages of using remote sensing in forestry inventories (Schreuder *et al.,* 1993):

- Satellite images offer a synoptic view of a study area;
- Satellite images cover large areas;
- Satellite image data can be digitally processed;
- Satellite image data is less expensive than aerial photographs.

Howard (1991) states that expectancy of Landsat applicability to forestry has been overstressed in relation to its operational role and that it is unlikely that satellite remote sensing will be able to provide the same quality and usable information as the aerial photographs.

Landsat satellite images analysis have successfully used for large area forest mapping and estimation of forest variables, forest changes detection over time and many other fields. However insufficient spatial resolution of Landsat images renders results less satisfactory for small areas inventories and for operative forest planning.

1.3 The use of remote sensing for forest inventories in Mozambique

Forest inventories in Mozambique are done for several purposes and scales. At national level, the first country based forest inventory was carried out in 1979/1980 pioneering the use of satellite images on national forest inventories. Forest inventory was based on the interpretation of MSS Landsat colour composite images from 1972-1973 covering the whole country (Malleux, 1980).

The objective of national forest inventories is to produce statistically unbiased, reliable forest resources information for large areas for strategic planning, primarily by decision makers (Katila, 2004). The sampling over nationwide forest areas and the natural variability of forests require large amount of field plots to obtain reliable forest information. This requirement has to be traded-off with restricted budget for forest inventories and the on-growing demand for more variables and information to be collected. Foresters desperately look for tools to enhance the precision of estimations and to reduce costs of forest inventories.

The satellite images, which cover vast areas, fitted to that expectation, specially for countrywise forest inventories. The MSS landsat images used on the first Mozambican national inventory where utilized mostly to enhance forest stratification and as a tool for sampling design. Despite the use of satellite images, the first national inventory results were considered with low degree of precision due to limited time and human resources as well as due to the use of medium resolution satellite data (Saket, 1995).

In Mozambique, satellite images where used as well, for forestry sector policy formulation on fuelwood and biomass energy subsector. 44 Landsat TM scenes from 1990-1991 covering the whole country where purchased to produce the country's biomass assessment (ETC, 1987). These same images where latter used to update the 1980's exploratory national inventory into two steps: (a) producing national forest map at the scale 1: 1 000 000 which did not gave enough details for development activities (Saket, 1995) and was followed up by (b) semi-detailed national forest types and land cover maps at 1:250 000 scale maps based on the 44 landsat MSS colour composed images (green, red and near infra-red bands).

These three works initiated the use of satellite images at nationwide level for policy and strategy formulation in the forestry sector. The inventory reports do not provide any information on the sampling error and estimations accuracy and results obtained where restricted by difficulties in obtaining field data because of the ongoing war and safety reasons. Some data such as tree species composition by forest types, diameter distribution for each species at district and provincial level , regeneration of natural forests and standing woody biomass where some of the non-existing information due to lack of comprehensive data gathering on the updating work of the national inventory (Saket, 1995). The same author recommended further detailed forest inventories based on representative sampling for the 10 provinces in the country to improve the existing information.

This recommendation highlighted the need to obtain more detailed information at provincial level since that information was needed to backup forest management decisions (forest concessions and investments on forest plantations). The first attempt to obtain forest data at provincial level was done on the immediate years after the country's independence in 1975 when two main large scale native forest inventories for strategic purposes where done in Niassa and Cabo Delgado provinces (Madebras *et al*., 1982, Silviconsult, 1984). Pancromatric black and white aerial photographs and topographic maps where the main tools for land use delineation, forest stratification and inventory planning on those days, but the war period restrict the use of the collected information and the inventory results.

After the country's peace agreement, strategic forest inventories at provincial level funded by World Bank where initiated in Sofala and Cabo Delgado Provinces with the broad objectives of identifying the most suitable areas for forest concessions and to allow the elaboration of management plan models for forest concession areas. Recently, this work has been expanded to Zambezia and Inhambane provinces with Finnish cooperation funds.

All provincial forest inventories where heavily based on field data collection and as an immediate work result, forest inventory sampling design, field plot sizes, form and types, plot location, establishment and measurement procedures were rapidly standardized and field manuals produced in the country. Overall forest volume sampling error has been achieved within the acceptable limits, although in some cases when individual forest types are considered the sampling error per forest type and species increases and overcome the given limits.

Satellite images, vegetation and topographic maps used during the mentioned inventories have provided the key information for forest stratification and sampling design. While for Cabo Delgado and Sofala forest stratification was based on the visual interpretation of image tones and subsequently manual delineation of polygons, for Zambézia province the first stratification was based on a pilot inventory spectral classes given by the normalized vegetation index. A further stratification was based on a false composite image visual interpretation (band 1, 2 and 4) of forest polygons coupled with land use and cover maps as ancillary information.

Continuous variables such as volumes per forest type, volumes per tree species, tree diameters and heights were gathered through field plots while forest type boundaries and area compiled and updated through satellite image visual interpretation. Total forest growing stock is estimated based on the forest type areas and mean volumes per forest types. This traditional procedure has been applied either for strategic large scale forest inventories such as the national inventories and provincial forest inventories of Sofala, Cabo Delgado and more recently Zambézia and Inhambane, either to local forests with envisage of management plans elaboration for community forestry and forest concession areas. The present study estimates Mocuba district volume based on the provincial forestry inventory plots as reference plots for expanding the information through all district by mean of satellite image spectral data and explores its applicability to district level forest estimations and management.

1.4 Integration of spatial data and forest inventories

Traditional forest stand parameters inventories are expensive and difficult to obtain specially if vast and heterogeneous forest areas are considered. With the availability of remote sensing and the possibility of acquiring an overview of large areas through images several approaches have been studied to combine field data with satellite data. The methods can be classified into two main approaches (Franklin, 2001):

a) regression approach

Predicts the selected field variable using the remote sensing data as the dependent variable. The main constraints of using regression analysis to estimate forest parameters are (Foody *et al*., 2001):

(i) Linear regression analysis assume that exists a linear relationship between the remotely sensed data and the forest variables to be estimated. In reality most relations observed may be curvilinear.

(ii) Multiple regressions assumes that the independent variables (that is, the data acquired by the sensors in different spectral wavebands) are uncorrelated. This assumption is seldom satisfied since data in different spectral wavebands are very strongly correlated.

b) *k*-NN approach or reference sample plots (RSP) approach

Non-parametric *k*-NN estimates forest parameters based on plot data and corresponding pixel values to estimate the parameters at all pixels, without assuming a model that relates spectral variables and field variables.

The use of predictive methods were no data distribution and models are assumed, such as the nearest neighbour algorithm, may offer an alternative to integrate spatial data and forest inventory ground data.

2. Problem definition

Provincial forest inventories in Mozambique are made for strategic purposes and global monitoring of forest resources. Satellite images are mainly used during the planning phase to delineate the main forest types and separate forests from non-forest areas in order to obtain the most adequate sampling design for vast areas.

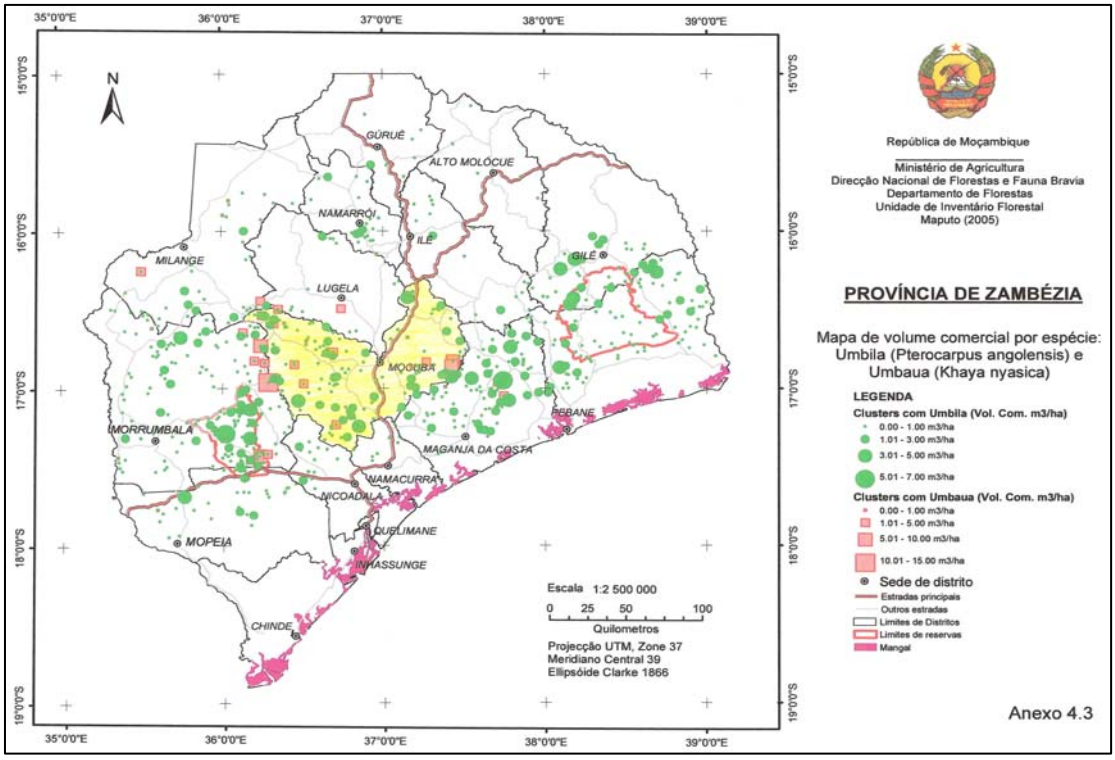
Traditionally, in Mozambique plotwise inventory has been used to derive the estimates for forest type level characterization.

The main problems of these forest inventory schemes can be summarized as:

1) Huge amount of information produced that requires long time for processing and report delivery. Zambézia forest inventory covered an area of about 4 million hectares. Data collection and processing required about 4 years to be completed and delivered to the general public and forest entrepreneurs. This is aggravated by the fact that the best forest areas are almost totally under logging activities with deficient registering (not georeferenced harvesting data). Forest inventory data produced becomes rapidly outdated.

2) The sampling design and intensity do not always allow the estimation of forest variables for a smaller units (such as districts or forest concessions) due to random distribution of samples and/or reduced sampling intensity in certain cases. In the recent Zambézia province forest inventory, results were calculated for aggregated district areas in order to obtain the desired level of accuracy of variables estimates. The applicability of provincial forest inventory data for forest management at local level (or even district level) may not be possible.

3) Mapping of continuous forest variables is not usually done due to the difficulties of spatializing plot level information to overall area based on sample plots. Therefore, the typical forest inventory main outputs is a forest type map complemented with figures and tables summarizing the characterization of forest types. Data as been considered difficult to understand and to relate to the ground information by the loggers and technicians. Figure 7 shows a map contained in the



inventory report (annex 4.3) which indicates the availability of volume for two selected commercial species: umbila (*Pterocarpus angolensis*) indicated by green circular dots and Umbaua (*Kaya nyasica*) indicated by pink square.

Estimation and mapping of forest variables using a non parametric method, such as the k-Nearest Neighbours might offer a possibility to extrapolate the data from field sample plots covered by a satellite image to the all area, covering low intensity sampled areas and even to non-sampled areas with a known error. The method has been classified as simple, straightforward and used in Finnish multi-source national forest inventories since 1990's. Its application has been mostly studied for boreal and temperate forests with successful results.

2.1 Research objective and assumptions

The present study has the following main objective :

- To estimate main forest characteristics at district level based on the forest inventory data collected at provincial level using the *k*-NN algorithm to expand the plot information to district.

When considering the possibility of propagating the forest plots inventory information collected, through the satellite image in order to estimate the forest characteristics within a defined geographic area, the main assumptions for using the *k*-NN algorithm are:

I. Forest inventory plots covered by a satellite image can be used as a ground truth plots for any subarea of the image (Tokola, 2000);

II. The pixel values depend only on the land cover and not on its geographic location (Killki and Paivinen, 1987). Since this assumption may not be true for the entire landsat image it is advisable to use a sub-population of Landsat scene or to include geographical distances in weights (Fazakas *et al.,* 1999);

III. There is a significant correlation among measured plot variables and the image feature since the quality of the estimation by the classifier algorithm relies on that (Franco-Lopez *et al*., 2001; Maselli *et al*., 2001);

IV. Reference plots cover all relevant variations in the study area (Franco-Lopez *et al*., 2001).

2.2 Research question

In Nordic homogeneous forests and plantations the k-NN extrapolating algorithm has shown successful results. In tropical and sub-tropical regions, characterized by a more complex and heterogeneous forest-stand structure with abundance of species, the method has not yet been spread and deeply studied as in the northern hemisphere forests. Its applicability to miombo forests, an ecosystem characterized by a certain degree of uniformity over vast areas in Africa is the object of this study. The research question is whether miombo forest variables can be predicted and mapped at a district level by means of the *k*-NN estimation method using Landsat data and regional inventory field samples, as reference plots.

2.3 Expected results and similar studies

Most of studies that attempted to estimate continuous forest variables by mean of high resolution images are recent (from 1970's) and were mostly carried out in boreal forests and plantations due to

their relatively simple stand structure and tree species (Mäkelä and Pekkarinen, 2004; Lu *et al*., 2004; Rajaniemi *et al.,* 2005). The method combine satellite image data with field data to estimate forest variables for each pixel of the image. After the extrapolation of forest variables for each pixel, the estimates for a defined forest area (such as forest stands) are calculated as an average of pixel estimates within the forest stand, compartment or administrative territory (Mäkelä and Pekkarinen, 2004). Although the *k*-NN method has been incorporated from 1990 onwards to the national forest inventory in Finland and has been applied successfully in boreal and temperate forest areas, in tropical forests there are just a few studies applying the method.

In southern Africa, an academic dissertation done by Sengupta (2003) for East Usambara Tanzanian tropical forests to detect areas with above-average three species richness used several predictive methods, including the non parametric *k*-NN. All predictive methods (logistic regression, multiple linear regression, discriminant analysis and k-NN) produced similar results accuracy. 69-72% of above average tree species richness areas were correctly predicted using the *k*-NN method.

Rajaniemi *et al.,*(2005) pioneered the use of *k*-NN in the Amazonian forests, to predict species richness (Pteridophyte and Melastomataceae species) using Landsat image and field observations for a rainforest study area in the Ecuadorian Amazonia. The study shows that although relative RMSE was quite high (66% for poor soils and both species predictions and 63% for rich soils and both species) all estimations were unbiased and the greatest absolute difference in rms error was 1.1 species for poor soils species group. The authors emphasized that the method showed encouraging results and practical applications of the *k*-NN method should be further investigated in the tropics, specially for those economically relevant forest characteristics, such as timber volume prediction and timber species.

From literature search it is expected that forest variables estimation error at pixel level is considerable in multisource forest inventories with high resolution satellite images (Curran and Williamson, 1985; Halme and Tomppo, 2001; Katila, 2004). In boreal forests, the predictions errors expressed with relative root mean square error (RMSE) are around 50-80% for total volume for pixel level predictions (Kilpeläinen and Tokola, 1999 citing Tokola *et al*., 1996; Katila, 2004). Errors in the volumes by tree species are higher than those for total volumes estimates.

The main reasons for high estimation errors at pixel level are: (1) image registration errors and (2) location errors of sample plots (Mäkelä and Pekkarinen, 2004).

Prediction accuracy for plot level slightly improves. Wallerman (2003) citing Tokola *et al*.,(1996), refer that for the boreal vegetation in Finland using relascope plots of a maximum radius of 12.54 m (about 500 m^2) the mean standard error of plot level predictions is 60% for total stem volume.

When estimates are done to stand level the prediction error reduces. In nordic country's conditions, characterized by very small size forest stands (about 1.5 ha) the estimation errors at stand level using the *k*-NN method varies from 26-68%. The lowest value (26%) was found for large and homogeneous pine stands variables estimations (Kilpeläinen and Tokola, 1999 quoting Hagner, 1998)

From the several studies carried out some results trends were extracted and summarized as follows:

- Prediction accuracy for a single pixel is often low (Curran and Williamson, 1985; Halme and Tomppo, 2001; Wallerman, 2003; Katila, 2004). The accuracy increases when single pixels are aggregated.

- The relative RMSE decreases rapidly when the area to predict increases, from single pixels to plots, from plots to forest stands or larger areas (Halme and Tomppo, 2001; Wallerman, 2003).

- *k*-NN estimates of forest stand borders pixels have larger bias than those inside the stand due to the presence of mixed pixels (Katila and Tomppo, 2001; Katila, 2004).

- The accuracy increases if auxiliary information is combined with satellite data (Holmgren *et al*., 2000; Tokola, 2000). Stratification of image and field plots significantly reduces de bias in volume estimates (Katila, 2004).

Although *k*-NN method is considered a feasible approach for forest and non forest land classes mapping production and area estimation (Franco-Lopez et al., 2001; McRoberts et al,, 2002; Haapanen *et al*., 2004), several authors have been conservative in raising the expectations regarding practical forest management planning due to large errors encountered at stand level (Kilpeläinen and Tokola, 1999; Holmström, 2001; Mäkelä and Pekkarinen, 2004).

3. Methodology background

3.1 The *k*-Nearest Neighbour method (*k*-NN)

The nearest neighbour classification procedure is one of the simplest and oldest methods for performing general non-parametric classification (Wagacha, 2003).

The classification is a two-step process (Andersen, 1998):

- training stage, which refers to the process of defining the criteria for which the patterns are recognized. It can be supervised or unsupervised. Unsupervised training is an automated computer process while in supervised training the user defines a set of classes which require *a-priori* knowledge about the area. This knowledge is obtained through reference or ground truth data. As a consequence, the supervised classifications are dependent on the quality and quantity of the reference data.
- classification stage, pixels or regions are assigned to classes based on a decision rule (distance or maximum likelihood). Once the classification is done the assessment of its accuracy is performed, as the final step of classification.

Katila (2004) refers that since E. Fix and J. L. Hodges in 1951 presented the nearest neighbour classifier, the algorithm has been extensively used in pattern recognition, that is, the assignment of a physical object to one or several specified categories. Pattern recognition methods are used in many fields. Among others: image preprocessing /segmentation, speech recognition, optical character recognition, fingerprint identification and seismic analysis (Gutierrez-Osuna, no date). The main attractiveness of nonparametric analysis relies on the fact that the functional form of the relationship between the response variable and the predictors variables do not follow a specified model creating an alternative to non-linear regression model when the theoretical model is not known or is difficult to fit (Andersen, 1998). Other strengths of the *k*-NN method for forest applications are (Rajaniemi *et al*., 2005; Katila, 2004): all interesting variables can be estimated at the same time and estimates preserve the original covariance structure of the field variables.

Of the non parametric regression methods the kernel and the *k* –Nearest Neighbour estimators are the most common in forestry applications (Katila, 2004). The kernel estimate is a weighted average of the response variable in a fixed neighbourhood , the *k*-NN estimate is a weighted average of the response variables in a varying neighbourhood defined by those that are nearest to the query point.

Considering a universe of data of known values (n_i also known as training dataset) and a data set x to be classified, for each x_i its k nearest neighbours are retrieved forming the neighbourhood of x_i. To decide the classification of x_i, it is used the majority of voting among the data records in its neighbourhood considering the distance between each data record known in the neighbourhood to x_i to predict or classify. The number of neighbours used to decide the classification or value of x_i varies from 1 to *n-1* and is known as the *k* value .

The main characteristics of *k*-NN method are (Gutierrez- Osuna, no date):

- It is a non-parametric approach, that is, distribution or density estimates are data driven and no particular distribution or model is assumed;
- Prediction is based on analogy, that is similarity between the query point and the other known k-neighbouring points;
- Similarity is expressed as the distance between the query point to a constant number the neighbouring points (*k*) with known variables values (also known as the training data set);

3.2 The use of *k*-NN in forest inventories

The use of the algorithm *k*-NN to assign forest inventory field data to satellite pixels has been firstly known as the reference sample plot method presented by Kilkki and Päivinen (1987) for area interpretation and thematic map production. It was firstly used in Finland and since 1990 it was incorporated into Finnish national forest inventory (Katila and Tomppo, 2001). In Finland, *k*-NN algorithm has been popularized in several forest fields with encouraging results: for basal–area diameter distribution (Haara *et al.,* 1997; Maltamo and Kangas, 1998), for generalizing sample tree information (Korhonen and Kangas, 1997) , for multisource forest inventory variables mapping and estimates (Kilkki and Päivinen, 1987; Tomppo *et al.,* 2001) and for wood procurement planning through the prediction of characteristics of a marked stand for harvesting (Tommola *et al*., 1999; Malinen*,* 2003)

The method was studied as well in Sweeden (Fazakas *et al*,. 1999), in European Mediterranean countries (Chirici *et al.,* 2003), in United States of America (Franco-Lopez *et al*., 2001, McRoberts *et al*., 2002, Haapanen *et al.,* 2004) and in other regions of the world (Rajaniemi *et al*., 2005, Tomppo *et al.*, 2001).

The main advantages of using the non-parametric *k*-NN method to estimate forest inventory variables relies on (1) no explicit model is needed in the estimation and (2) all measured variables can be estimated at the same time. This last mentioned advantage is specially important if it is considered that large scale forest inventories (national or regional inventories) involve high number of variables (Katila and Tomppo, 2001). This is even stressed with the present trend in forestry for measuring other products that just timber increasing the number of variables that were traditionally measured on forest inventories. Estimates of all forest variables measured are needed making the inventory processing stage a tedious work.

A third advantage of the *k*-NN method for forest inventory processing relies on the fact that covariance structure of the variables are preserved, since the same neighbours are involved for a number of forest variables estimations and the quality of estimations is influenced only by the correlation among the variables and the image features (Halme and Tomppo, 2001 citing Tomppo, 1991; Franco-Lopez *et al.,* 2001).

Several authors mention other additional advantages of using *k*-NN algorithm to estimate forest inventories variables with multi-source forest inventories (Tuominen and Poso, 2001; Franco-Lopez *et al*., 2001): its flexibility, simplicity and straightforward applicability in different forest conditions and with different remote sensing materials.

Although *k*-NN method has been very promising for extrapolate forest inventory variables throughout the landscape there is room for improvement and research is needed (Franco-Lopez *et al.,* 2001). One limitation of *k*-NN method is given by representativeness of the field observations: spectral values on the field data should cover the full range of spectral values in the target area (Rajaniemi *et al*, 2005). Another limitation is related to the quality of estimation analysis, which has not been yet explored in depth . There is a need for research on methods to evaluate estimation errors and to improve the accuracy of estimations through stratification, research about ancillary forces affecting local response, research on the best way to use the ancillary information and plot location accuracy research (Franco-Lopez *et al.,*2001).

In tropical countries, the use of *k*-NN for forest estimations is recently being studied and its applicability within an heterogeneous forests conditions it is not yet fully understood. Rajiniemi *et al.* (2005) used *k* equal to 3 for estimating and mapping species richness in Ecuador Amazonian rainforests and results showed that *k*-NN method could predict field variables using Landsat TM data.

3.3 How to improve *k*-NN accuracy of forest estimations

Katila (2004) studied the source of errors in the Finnish multi-source forest inventory and found the following categories of sources of errors: (i) aspects related to inherent characteristics of sensors and forests, that is the low dynamic range of spectral bands on high resolution optical data, (2) aspects related to sampling design, being the small size of the national forest inventory field plots (about 490 m^2/ 12,5 m radius) compared to the pixel size in image data and (3) aspects related on accuracy of multi-source data matching, being the most important ones locational errors in the image and field plot data.

Several authors describe and use several techniques to minimize the different sources of errors. Errors derived by the first mentioned aspect, that is, inherent technology characteristics cannot be easily changed and therefore are not here considered.

When combining the ground plots and the satellite image data there are several aspects that contribute to improve the accuracy of forest variables estimates. They are (Katila and Tomppo, 2001; Halme and Tomppo, 2001; Katila, 2004):

a) Sufficient field sample size and sampling design

The number of samples varies from different studies and it is difficult to establish a rule of thumb for the number of field plots required. Most of studies use national forest inventory plots which provide an immense amount of field data, as high as 10.000 plots per image. Haapanen *et al.,*(2003), citing Nilsson (1997) mention that at least 500 forested field sample plots should be used to estimate forest volumes for boreal forests. Tomppo *et al.*(2002) indicate that the limiting factor of k-NN method for estimating biomass and volume in boreal forests is the need for considerable amount of ground data. The author indicates that for using a full Landsat TM scene at least 1000, preferably 5000, georeferenced field plots are needed.

The driving force on sampling intensity selection is to have enough field samples to cover all forest estimation parameters variation (tree sizes, forest types, forest volumes, and so on) (Franco-Lopez *et al.,* 2001).

Those errors occurring due to sampling design are restricted by the fact that field data used is usually derived from a design already made, that is the existing national/provincial forest inventory field plots. Sampling design cannot be change to fulfill the requirements of a multi-source non-parametric forest variables estimation model.

Forest inventories in Mozambique use cluster sampling for practical reasons due to more efficient time ratio between plot measurement and walking. The main problem of a cluster design when compared to regular layout of plots is the probable spatial correlation of variables between the neighboring plots, that is, within clusters (Tomppo *et al.,* 2001). If distance between plots in increased, the correlation between plots within a cluster is reduced. But this distance cannot be infinitely increased since the plot layout and design is also determined by one-day field work team productivity. A distance of 200 meters is usually used between plots within 4 plots per cluster design in the Mozambique native forests.

Haapanen *et al.*(2003) detected that in a cluster design sampling there is a tendency for the target pixel to select nearest neighbors within its own plot and cluster. If distance search radius was limited around the target plot the probability of choosing nearest neighbors from the same target plot increases significantly, when compared to unconstrained search.

b). Satellite images are accurately rectified

Accuracy of estimations are also dependent on the accuracy of pixel location in relation to the real position in the ground.

c) Coordinates of the field plots are accurately known and positioned at the image

Pixel values extracted from the ground plots are related through *k*-NN method to propagate the estimation of defined variables to overall or parts of the satellite image scene. Therefore, pixel-level estimates are significantly affected by field plot dislocations and rectification errors (Halme and Tomppo, 2001; Katila and Tomppo, 2001; Katila, 2004). With the use of global positioning system (GPS) the extraction coordinates of ground plots become more easy and accurate. The gained accuracy offered by Global Positioning Systems is likely to improve the inferences between pixel data and ground data and therefore the corresponding extrapolations and predictions (Halme and Tomppo, 2001). Although plot dislocation reduces with the use of the Global Positioning System (GPS), there is always an inherent instrumental accuracy combined with map errors and satellite image rectification errors, that causes field plots dislocation. In Finland, the total assignment error of field plot to a satellite image pixel with a pixel size of 25 m, showed a variation from 0 to 70 meters (Halme and Tomppo, 2001).

To minimize the influence of plot dislocation on the pixel level estimations, a selection of a high value of *k* reduces the pixel level variation of the estimated variables, controlling in some way the errors of plot dislocation (Tomppo *et al.,* 2001). The use of several window sizes somehow reduces the plot dislocation errors.

d) *k*-NN parameters calibration:

When the *k*-NN method is applied to multisource national forest inventories, which combine the use of satellite images, digital maps and field data plots, a set of parameters must be selected or calibrated to our own dataset to minimize the overall *k*-NN estimation error or any other defined criterion (Franco-Lopez *et al*., 2001; Katila and Tomppo, 2001; McRoberts *et al*., 2002; Malinen, 2003; Katila, 2004).

When sampling design is fixed and cannot be changed the *k*-NN parameters calibration constitutes the only possible action to reduce inaccuracies from *k*-NN estimations and therefore this aspect is detailed in point 3.4.

3.4 *K*-NN parameters selection/ calibration in multisource forest inventories

K-NN parameters can be classified into two main groups (Katila, 2004):

I) Parameters related to *k*-NN local tuning:

1) The distance measure.
2) The value of *k* or number of nearest neighbours.
3) The image features and band selection.

II) Parameters related to selection of training data:

4) Stratification of the image and field plots.
5) Geographical reference area from which the nearest neighbours are selected.

The appropriateness of parameters selected or *k*-NN calibration is guided by an optimization objective (McRoberts *et al.*, 2002; Katila, 2004). The most common guiding criterion when estimating volumes is the minimization of the overall *k*-NN estimation error usually assessed by the

leave-one-out criteria (cross validation). But other criteria can be included when calibration of *k*-NN parameters is performed (Katila, 2004). For example, the tuning of *k*-NN parameters is guided by retention of some of the variation of the original ground data in the spatial variation of estimates (Katila, 2004). This criteria specially important when mapping the vegetation classes is the primary goal .

3.4.1 Choosing a similarity measure: distance

In one dimensional space (D1) the distance between two points is characterized by one parameter. In multidimensional space (D>1) there is no unique mathematical way to define distance between two points, since it represents a measure (d) between to points *x* and *y* that has to satisfy the following properties:

i. $d(x,y) \geq 0$ if $x \neq y$ (*non-negativity property*)
Distance from two different points (*x,y*) is always positive.

ii. $d(x,x)=0$, and $d(x,y)=0$, that is $x=y$ (*self distance property*)
Distance from any point to itself is zero;

iii. $d(x,y)=d(y,x)$ (*symmetry property*)
It is equally distant from *x* to *y* as from *y* to *x*;

iv. $d(x,y) \leq d(x,z)+d(z,y)$ (*triangle inequality property*)
It can never be shorter to go to an intermediate point than to go directly ;

When using *k*-NN three types of distance calculations are usually selected:

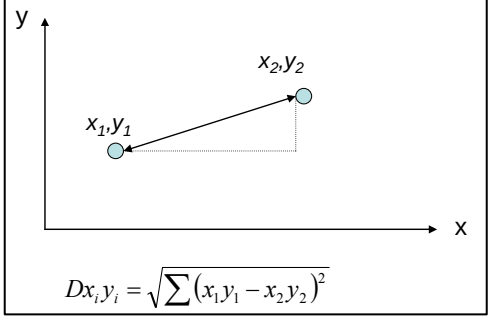
- Euclidean distance,
- Manhattan distance and
- Mahalanobis distance.

The selection of distance type influences the " shape" of neighbourhood that will be applied.

<u>Euclidean distance</u>

The Euclidean distance D is the shortest distance (straight line) between two points in a plane. Its value is found by resolving Pythagoras' theorem for a right angled triangle formed from two points in cartesian coordinate system (bi-dimensional) (Figure 8).

Figure 8 - Euclidean distance



Such distance is defined as *Euclidean metric* or L2 metric distance. The Euclidean distance defines a neighbourhood with a circular shape around the point to be predicted if the points have all the same distance to the origin. In digital images the elements of *xi* and y_i are columns and rows and due to its simple features it is widely used in *k*-NN analysis.

When applying the Euclidean distance to between neighbours on satellite images, the distance between pixel *p* (to be classified) and pixel p_i (training pixel with ground data known) , the general expression for the Euclidean distance is (Franco-Lopez *et al*., 2001):

$$dp(pi) = \sqrt{\sum_{j=1}^{nf} \left(x_{p,j} - x_{(p),j}\right)^2} \qquad \text{[eq.6]}$$

where $x_{p,j}$ is the digital number for the feature *j* and *nf* is the number of features in spectral space.

Manhattan distance

Manhattan distance D is defined as the rectilinear route measured along parallels to *x* and *y* axes as shown in figure 9.

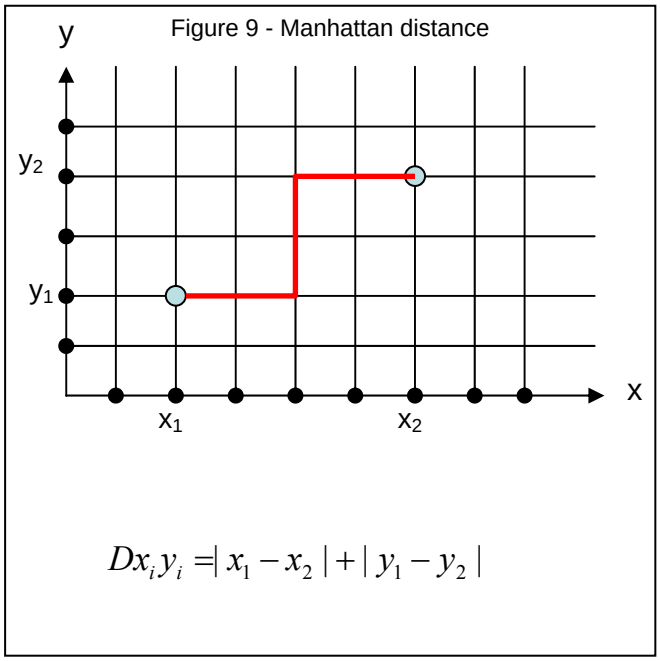


Figure 9 - Manhattan distance

It is called the Manhattan metric because it is the distance a car would drive in a city laid out in square blocks and also known as L1.

Mahalanobis distance

Mahalanobis distance was invented in 1936 by Prasanta Chandra Mahalanobis that pioneered the consideration of correlations between variables and by which different patterns could be identify in respect to the reference point. It differs from the Euclidean distance because it takes into account the distribution of points (correlations) and that constitutes the main strengthness of this distance.

The Mahalanobis distance assumes a normal distribution of the band histograms and uses the means and the covariance for each variable. Therefore, when determining the distance of a variable from a centre, components with high variability should receive less weight than the components with low variability.

The Mahalanobis distance between two points $x=(x_1,....,x_p)^t$ and $y=(y_1,....,y_p)^t$ in the p-dimensional space $\mathbb{R}^P$ is defined as:

$$ds(x,y) = \sqrt{(x-y)^t S^{-1} (x-y)} \qquad \text{[eq.7]}$$

where:
S^{-1} is the inverse covariance matrix.

If all points have the same distance to the origin it is obtained an ellipsoid .

Distance weighting or neighbours weighting functions

The *k*-NN classification technique is intuitively based on the assumption that objects close in distance to the query pixel *p* are more potentially similar and should be given greater weight. Based on this assumption each neighbour can be weighted (w_i) according to the general function for weighting distances (Franco-Lopez *et al.*, 2001):

$$w_{(pi)p} = \frac{\frac{1}{d^t{}_{(pi)p}}}{\sum_{j=1}^{k} \frac{1}{d^t{}_{(pi)p}}} \qquad \text{[eq.8]}$$

The spectral distance d_p is computed from the pixel *p* to be classified to each pixel p_i with the ground measurements known. *t* can varies from 0 to 2, obtaining a neighbour weighting function:
(a) inversely proportional to the distance (t=1);
(b) inversely proportional to the square distance (t=2) ;
(c) equal or nor weight (t=0).

Franco-Lopez *et al*. (2001) compared the three options to weight neighbours for St. Louis county - USA and concluded that to take full advantage from the weighting inverse and squared inverse functions it is needed to have some distance between an identifiable close neighbour and the next one. This was not the case in the study since the data was collected from plots regularly spaced on the landscape but in the case in cluster design forest inventories applied in Nordic countries where substantial distance exists between clusters, weighting distances presents advantages.

The inverse distance weighting method (inversely proportional or squared) has been widely used in *k*-NN analysis in many different studies from Finland and Sweden (Katila and Tomppo, 2000; Tomppo *et al.,* 2002; Reese *et al.,* 2002; Nilsson, 2002). The *k*-NN software utilized in this study uses the inverse distance weighting method, that is those training plots closer to the query point are considered first.

3.4.2 Choosing the value of *k*

The choice of *k* values (number of neighbours to be considered as training data) is one of the most important factors when using *k*-NN and can strongly influence the quality of predictions (McRoberts *et a*l., 2002).

When only 1 nearest neighbour is used (*k* =1) it is obtained an unstable classifier expressing high variance and highly sensitive to data used, that is, it incorporates all the variability that exists in the observations (McRoberts *et a*l., 2002). Franco-Lopez *et al*. (2001) suggest that for mapping purposes a *k* value equals to 1 should be used because it retains the variation of original data.

Classification can be improved by increasing the value of *k,* but the number of neighbours cannot be indefinitely increased because there is a trade-off between variance and bias. When the value of *k* is increased the variance is reduced but increase in bias can be obtained due to averaging values, resulting in a tendency of underestimating high values and overestimating low values (averaging) in the *k*-NN method (Tuominen and Poso, 2001).

Thus, *k* value should be set to a value that is large enough to minimize the probability of misclassification and small enough to minimize the bias.

In northern Europe countries and in USA some studies have been conducted to determine the optimal *k* value and the variables affecting it (Franco-Lopez *et al*., 2001; Tokola *et al.,*1996). Franco-Lopez *et al.* (2001) found that in *k*-NN estimations for basal area the values of RMSE decreased about 14% when the value of k was increased from one to five, but when the *k* was equal to 9 or above the marginal increase in precision was smaller than 0.5%. Other authors (Tokola *et al*., 1996; Nilsson, 2002) reported as well a stability point between 10 and 15 neighbours.

In practice, a *k* value higher than 1 compensates for noisy in real data caused by georeference errors, atmosphere scattering, radiometric resolution of the sensor, incomplete sampling of the spectral variations in the image by field data, field plot dislocations and differences in size and shapes of pixels and field plots (Franco-Lopez *et al*., 2001; Katila and Tomppo, 2001; Halme and Tomppo, 2001).

Tomppo and Katila (2001) express that the optimal number of neighbours (*k* value) is affected by:
(a) the layout of field plots,
(b) the size of the field plot compared to pixel size,
(c) variation of field variables, given by the density of training data in spectral space.

Other authors expressed that the optimal number of neighbours was dependent on: (a) data and (b) goals of the estimation and survey (Franco-Lopez *et a*l., 2001; Reese *et al*., 2002 citing Hagner *et al*., 2002).

Katila (2004) summarizes the various values of *k* used in boreal forests by several authors: 1, 5 to 10 and 10 to 15. He found that for total volume estimates the *k* value varied from 7-11 considering the criteria of minimum relative RMSE decrease of 0.5% between *k* and *k*+1 sought from a window ranging from *k*+1 to *k*+5.

McRoberts *et al*. (2002) refers that a resulting *k* should not be straight forward extrapolated to other studies with different datasets since is affected by several factors. He also pointed out that an objective criteria for selecting the optimal *k* must be used in order to select the most suitable *k* value. The most common used criteria for *k* selection is to choose the one that minimizes the misclassification error using the cross validation evaluation parameter (root mean standard error) to judge this criteria. To select the appropriate value of *k* cross validation classification accuracy (RMSE) is plotted against k. The *k* value that register the smallest RMSE is the most appropriate to our dataset.

As a conclusion, when selecting the *k*-value three aspects are important (McRoberts *et al*., 2002): (1) use an objective selection criterion, (2) user should report the criterion and (3) users should justify their selection.

Validation and cross –validation

Validation is the process of determining if a mathematical model of a physical event represents the actual event with sufficient accuracy (Babuska *et al*., 2003). Ideally data should be collected to create a forecast and new data should be collected to verify the forecast. When this is not possible, validation estimates the generalization error of the model based on resampling or "data split-method ", where only a single data subset (validation dataset) is used to estimate the generalization error.

In cross validation differs from validation because the generalization error is estimated with different *k* data sub-sets. It is an evaluation technique that mimics the use of independent data by omitting training sample one by one (Franco-Lopez *et al*., 2001).

The assumptions for cross validation are (Fowler, 2004):

a. The testing and training sets are independent.
b. Observations in the testing and training sets are from the same distribution.
c. Testing sets observations are accurate and unbiased.

Error is the difference between the exact solution of the mathematical problem and its approximation. The root mean sampling error (RMSE) is the most used for continuous variables

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad \text{[eq.9]}$$

where y_i, i=1,…,n are the values of variables in the training dataset and $\hat{y}$ is the estimated value through the k-NN method.

The relative RMSE (RMSEr) can be calculate as a proportion of the mean estimated value (Mäkelä and Pekkarinen, 2004):

$$RMSE_r = \frac{RMSE}{\bar{\hat{y}}} \qquad \text{[eq.10]}$$

The reliability of estimates can be also measured by means of bias ($\overline{bias}$) according to the equation:

$$\overline{bias} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} \qquad \text{[eq.11]}$$

where $\hat{y}$ is the estimated value by *k*-NN procedure, y_i the observed value on the validation field plots and *n* the number of observations. The relative bias ($Bias_r$) is calculated as a proportion of the mean estimated value $(\bar{\hat{y}})$ through the equation:

$$Bias_r = \frac{bias}{\bar{\hat{y}}} \qquad \text{[eq.12]}$$

To test whether any observed bias is statistically significant from zero the t-student test statistics can be used by assuming independence and normal distribution of errors through the equation of standard error of bias ($S(\overline{bias})$) (Rajaniemi *et al*., 2005):

$$S(\overline{bias}) = \frac{SD_{bias}}{\sqrt{n}} \qquad \text{[eq.13]}$$

Where SD_{bias} is the standard deviation of differences between the validation plot value and the pixel estimates through *k*-NN (that is, bias) and *n* is the number of observations.

The standard error of the bias indicates if the bias deviates significantly from zero. When deviations are greater than $2.s(\overline{bias})$ the bias it is considered to be statistically significant (Katila and Tomppo, 2001). Other way to evaluate if bias deviates significantly from zero is to calculate the *t* bias through the equation (Tokola, 2000):

$$t_{bias} = \frac{Bias}{\frac{SD_{bias}}{\sqrt{n}}} \qquad \text{[eq.14]}$$

Bias is significantly different from zero when t_{bias} greater than the t student tabulated (usually 1,96 considering a larger number of plots, at 0.05 probability level).

Possible methods to perform both validation and cross validation of a *k*-NN prediction model are (Katila and Tomppo, 2001):

a) data splitting or "hold out": in this approach, a subset of the available data is set aside to serve as a "validation" set. Dataset is usually randomly divided into two equal datasets (one as training data and one as validation data). This method is often sufficient when abundant data are available.

b) Leave one out cross validation: If the amount of data available is not adequate to provide appropriate amount for training and testing data, the division of data available into two equal size datasets might not be possible. It is also known that hold out validation is sensitive to the size of validation set (and hence to the training set). One way to reduce the sensitivity of validation methods on the choice of splits is to repeat the validation procedure a number of times and average the results.

   Leaving-one-out is a straightforward technique for estimating classifier error. The leave one out method apply the prediction *k*-NN algorithm to the data set omitting one sample by one at a time (i^{-1}) and summarizes the error that the *k*-NN prediction makes when it predicts the omitted unit.

The leave-one-out error rate estimator is considered nearly unbiased estimator of the true error rate of a classifier, although can be highly variable for small samples where the variance of the error estimates dominates (Franco-Lopez et al., 2001 citing Efron and Tibshirani, 1997). It has been initially mostly applied to small sample datasets due to its almost unbiased estimation of error and the fact that all the cases in the available sample are used for testing and almost all the cases are also used for training the classifier. Therefore the computational costs were initially considered expensive for large sample datasets. With the development of computers memory and speed the use of the technique has been expanded to large datasets and used with satellite image and forest inventories datasets and many authors use the cross validation method to determine the accuracy of *k*-NN based estimations (Halme and Tomppo, 2001; Haapanen *et al*.,2004).

3.4.3 Image features and band selection

Landsat 7 has a multispectral sensor, registering, information from different wavelengths of light, either visible wavelengths, infrared and thermal wavelengths. Band selection affects on which features can be best seen in a image. Table 1 describes the seven Landsat TM bands and how they are tailored to detect different earth features:

Table 1 - Landsat band and features

Band	features
Band 1 (0.45-0.52 µm, blue-green)	It is used to monitor sediment in water, mapping coral reefs and water depth. It is also useful for distinguishing deciduous from coniferous plants, as well as for distinguishing soil from vegetation.
Band 2 (0.52-0.60 µm, green)	Matches the wavelength of green vegetation. Used to assess the health and vigor plants.
Band 3 (0.63-0.69 µm, red)	It also called the chlorophyll absorption band and is useful to distinguish vegetation and soil, to discriminate between vegetation types and to monitoring vegetation health.
Band 4 (0.76-0.90 µm, near infrared)	Water absorbs barely all light and water bodies appear very dark. This enhances the contrast between water and land interface. It is useful for mapping shorelines and determining biomass content.
Band 5 (1.55-1.75 µm, mid-infrared)	Band very sensitive to moisture and used to monitor vegetation and soil moisture. Also used to differentiate between clouds and snow.
Band 6 (10.40-12.50 µm, thermal infrared):	As a thermal band measures the surface temperature. It is primarily used for geological applications and to differentiate clouds from bright soils, since clouds tend to be very cold.
Band 7 (2.08-2.35 µm mid-infrared)	Used for vegetation moisture detection although band 5 is most preferred for that and for soil and geology mapping.

In multispectral images band combination of three different bands produce different types of image composites. The specific bands used in three-band composites are usually identified by giving the band numbers used for red, green, and blue in this specific order. The selection of appropriate band combination for an RGB image depends on the purpose and objectives of the work. Table 2 describes the main combinations used with Landsat images.

Table 2 - Main image composites types

Band / monitor colors	Composite image type	description	Most used purposes
3 2 1 / R G B 3= red band 2= green band 1= blue band	Natural / true colors	True-color composite images match the spectral range of vision for the human eye, so these images appear to be close to what we would expect to see in a normal color photograph. Active vegetation appears green, bare soil and fallow fields are brown, and urban structures are white. Clean water bodies appear black. River displays as muddy brown color	The image tend to have low contrast and somewhat hazy in appearance. This is because blue light is more susceptible than other bandwidths to scattering by the atmosphere.
4 3 2 / R G B 4= near infrared 3=red band 2= green band	Near infrared composite also known as " standard false color composite"	The visible blue band is replaced by the Near Infrared (NIR) band in this composite. Active vegetation is vividly depicted in as varying tones of red-pink. Bare soils and fallow fields are green and urban structures are bluish-white. Water, which absorbs nearly all of the NIR energy that reaches its surface, appears very dark, nearly black. Rivers displays as green brown color.	432 composite is very used to determine the extent of vegetation and to classify different vegetation types as seen from space. On the other hand this type of composite imagery would not be useful for studying underwater features.
5 4 3 / R G B 5= mid infrared 4= near infrared 3=red band	Special color composite	Active vegetation appears green and stressed vegetation appears brownish, facilitating differentiation. bare soil and fallow fields are red-brown, and urban structures are purple. Most water bodies are black.	Depicts water/land interface and is very similar to true-color composite of earth's surface. Used for coastline studies. The 543 is useful for forest studies. 543 composites looses information from soil types which are less separable because the image looses information from Band 2 where soils have a high reflectance.
7 4 3 / R G B 7=mid infrared 4= near infrared 3=red band Or 7 4 2 / R G B	Shortwave infrared composite	A Short-wave Infrared composite contains at least one band in the short-wave infrared (SWIR) portion of the electromagnetic spectrum. The other bands used can vary depending on the use of the composite data. The SWIR portion of the spectrum detects primarily the moisture content of the feature. 7 4 2 image consists of green, near-infrared and mid-infrared light portrayed in a false-color manner. Active vegetation appears green, bare soil and fallow fields are red-brown, and urban structures are purple. Most water bodies are black, and rivers appears dark blue	SWIR composite image are used to detect vegetation under stress (drought, pests) and in detecting soil types and soil disturbance since moisture is an important characteristic of soil structure.

Choosing bands aim to isolate those that are most useful in portraying the essential elements of an image, reducing the number of bands that must be analyzed and consequently reducing the computational demand (Campbell, 2002). Variance–covariance matrix shows interrelationships between pair of bands. High correlation (above 0.9) between pair of bands means that the values in the

two channels are closely related and they tend to duplicate information in the other. Feature selection attempts to identify such duplication so that the data is analyzed with the minimum channels or bands that provide the maximum information.

Reese *et al*. (2002) refer that many studies have investigated the relationship between wood volume and satellite spectral reflectance. Landsat image show a negative correlation with wood volume except for near-infrared band which studies show various correlations.

Pearson correlation coefficient (r) is used as a measure of the degree of linear relationship between two variables and is defined as:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} \qquad \text{[eq.15]}$$

Where x and y are the individual measurements of the variables.

Correlation coefficient varies from –1 to +1. Negative sign indicates a negative linear relationship while the positive sign indicates a positive relationship between the variables. Correlation coefficient values, given enough number of samples (n >30), are classified as:

- 1 to – 0,7 – strong negative association between variables
+0,7 to 1,0 - strong positive association
- 0,7 to -0,3 - weak negative association
+0,3 to +0,7 -weak positive association
-0,3 to +0,3 - little or no association

Pearson correlation coefficient work best when variables are approximately normally distributed and have no outliers. The main limitations of the Pearson's correlation coefficient are:

- It requires interval scale data – it therefore cannot be used with ordinal or nominal data forms.
- It is parametric - specifically, it assumes that both variables are normally distributed.
- It assumes that the relationship between the two variables is linear and consistent in direction.

Regression analysis is another method that provide information about a set of independent variables from where we want to extract the best subset to build a model. The stepwise regression is a semi-automated process of model construction by successively adding or removing variables (forward and backward process respectively) based only on the *t*-statistics provided by statgraphics software. For each variable currently used in the model it computes the *t*-statistics for its estimated coefficient, squares it and reports this as "F to remove" statistics of that variable. For each variable not in the model, it computes the t-statistics that its coefficient would be if it were next variable added, squares it, and report as the "F to enter" statistics. In the next step the programme automatically enters the variable with highest "F—to-enter" statistics and removes the variables with the lowest " F-to-remove" considering a F threshold values. The threshold values are established considering that F-statistics are squares of corresponding t-statistics. Therefore an F-statistics of 4 is equivalent to a t-statistics of 2, which is the rule of thumb value for 5% significance level. A threshold of 3 or 3.5 allow more variables to enter in the model and is set by default. This do not represent a problem since the stepwise algorithm does one variable at a time, and if more variables are allowed to enter they can be manually removed after results interpretation.

3.4.4 Stratification of image and ancillary data

Ancillary data can be defined as data acquired by means other than remote sensing that are used to assist in the classification or analysis of remotely sensed data (Campbell, 2002). Ancillary data is

incorporated into the analysis of remote sensed data and the primary requirements for such data are (Campbell, 2002):

- Ancillary data is available in digital form.
- Ancillary data pertain to the problem to be analysed.
- Ancillary data is compatible with the remote sensed data.

When working with vegetation and forest classes, elevation data is crucial for mapping vegetation patterns with digital data since vegetation patterns are closely related to topography. Topography is digitally represented in Digital Elevation Models (DEMs) which represent the shape of earth's surface and each pixel represents an elevation measurement, rather than a brightness value.

Data for DEMs can be compiled using one of the three alternatives (Campbell, 2002):

- Contours from maps can be digitized, converted to vector files and tagged with elevation values producing an elevation map;
- Photogrammetric methods derive the elevation map from stereoscopic images;
- Elevation maps are constructed based on survey data but they are very expensive and time consuming.

DEM's offer the possibility to incorporate, manipulate, classify and display elevation data to improve the classification task. Digital elevation models facilitate the image stratification into ways: (1) "by superimposing" since topographic data, such as elevation and slope can be hardly extracted from the spectral information in the satellite image, and (2) "by adding" when digital elevation model is incorporated as an additional band of data for the classification procedure.

3.4.5 Choosing the geographical reference area

When applying the *k*-NN method to expand the sample plots values to the overall area covered by the image it is assumed that the reference plots cover variations of both horizontal and vertical directions representative of the area to which the data will be estimated.

In Mozambique forests are mostly lowland forests (Saket, 1994) showing an increase in stock volumes per unit area as well as average tree diameters and height from the south to the north direction of the country and from the lowlands of coastal areas to the mountainous inland regions.

A satellite image scene (Landsat 5 or Landsat TM 7) covers an area of 183 x 172 km (approximately 3,2 million hectares) of forests and other land use classes. In such great magnitude, forests can show gradual changes in vegetation structure and in other forest variables. The relationship between growing stock and image features may also vary because of changes in the average structure of the growing stock (Katila and Tomppo, 2001). The larger the geographical area (both horizontal and vertical) covered by the training data set, the better it covers the true variation of the values in the field. When using the *k*-NN to expand the information from the reference plots to the overall pixel image it is assumed that the non-sampled subareas of the image have the same characteristics as the larger area covered by the image and the reference sample plots or training plots (Tokola, 2000).

3.5 Algorithm appropriateness

An algorithm is a procedure or a set of rules that we use to solve a particular problem (Wainwright and Mulligan, 2004). They are a tool on the modelling process, that is the representation in abstract form of a real system, and there are many ways and equations to represent a environmental system and its components in order to solve a particular problem.

The appropriateness of an algorithm is mainly judged (1) in terms of the process representation and (2) in terms of computational resources required (Wainwright and Mulligan, 2004).

In process representations, assumptions are usually needed and the value of the model depends on the validity and the scope of those assumptions. Assumptions must be important for the purpose for which the algorithm is intended and must be clearly stated. Other important aspect of process representation is related with the simplicity of the model. Variables used must be always relevant to the model and the greater the number of variables used the greater the likelihood that they will be cross-correlated. Perrin *et al.* (2001) show that simple models can achieve as much as the same level of performance of more complex models.

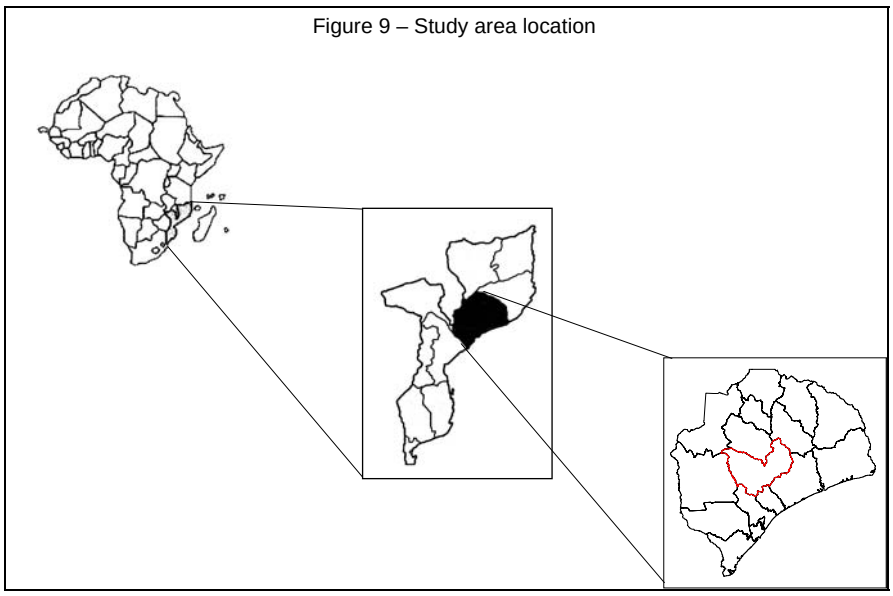
In terms of computational requirements it can be mainly considered that large amounts of spatially distributed data is used (Landsat images and digital elevation models) in environmental modelling which requires large amount of virtual or volatile memory to perform calculations (Wainwright and Mulligan, 2004). Thus, criteria like virtual memory needed while the model is running and time required to produce outputs can be considered when choosing the *k*-NN algorithm.

4. Materials and methods

4.1 Location of study area

Mocuba district is located in Zambezia province (central Mozambique and north of Zambezi river) and was selected as the reference area for *k*-NN analysis (Figure 9). Located in lower Zambézia province the district presents two seasons: dry winter from April to October and rainy season from November to March. Total annual precipitation is 1200 mm (average for the period from 1916 to 1968). Mean temperature varies from 22 ^{o}C during dry winter months to 27 ^{o}C for the wet and hot period. Mocuba district topography is characterized by an almost flat to ondulating topography.

Figure 9 – Study area location



The most recent and available data on forest resources of Mocuba district provides two forest figures for the district: 499 100 hectares inventoried by the forest inventory unit (Cuambe *et al.,* 2005) and 500 896 hectares derived from the forest map.

Mocuba district was selected as the reference area on this study due to the following main reasons:

- number of forest inventories sample plots available;
- availability of forest inventory figures and sampling error for the district that could be used as reference and comparison material;
- ondulating and almost flat topography;
- relative importance in terms of forest resources and wood industry supply area.

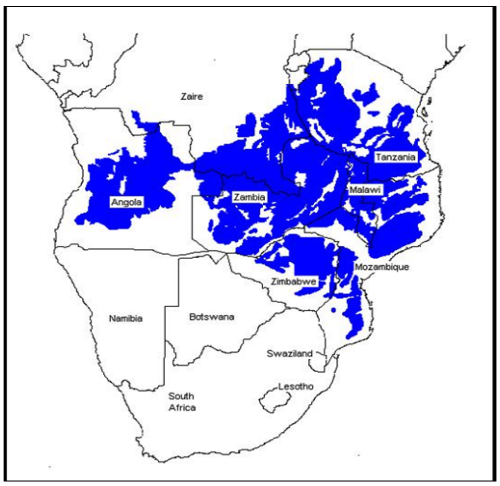
4.2 Forest description and characteristics

Miombo woodlands ecology has been studied by several authors and is described has the most extensive dry forest formation in Africa (Frost, 1996; Desanker *et al*., 1997; Kowero *et al*., 2003). It covers about 2.7 million Km^2 across southern sub-humid tropical belt of Africa over seven African countries, namely: Angola, Malawi, Mozambique, Tanzania, Zaire, Zambia and Zimbabwe (figure 10). It as been considered the largest more-or-less contiguous tropical woodlands and dry forests in the world (Desanker *et al*., 1997).

Although miombo woodlands has been classed in the past as savanna and can be considerably heterogeneous in terms of tree height, canopy cover and herbaceous structure it can differentiated

from other african forest formations by the dominance of trees in the subfamily Caesalpinioideae, being *Brachystegia*, *Julbernardia* and *Isoberlinia*. World Wide Fund for Nature (WWF-SARPO, 2001) has defined the miombo forests as an ecoregion in Africa dominated by "pure" miombo forests and including other related dry woodlands such as *Colophospermum* forests (mopane), *Acacia/Combretum* association forests, *Burkea/Terminalia* association forests and others.

Figure 10 - Miombo distribution



Fonte: Desanker, P.V. and Pretince, I.C. 1994. Miombo – a vegetation dynamics model for the miombo woodlands of Zambezian Africa. *Forest Ecology and Management* 69.

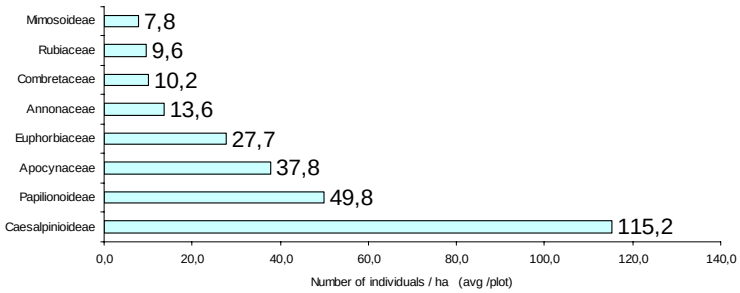
In Mozambique, miombo is the most common forest type and is mostly predominant in the country's northern provinces. Miombo forests occur from Rovuma river in the north to Limpopo river in the south. In northern areas it is almost a contiguous vegetation type and in certain areas is fragmented with other vegetation types (coastal forests, mopane and other dryland vegetation types). Due to its vast coverage, miombo forests occur over various types of topography, soils and land use areas in the country. The floristic uniformity and monotony of vegetation is the main feature of miombo forests and overrides the diversity and heterogeneity caused by variations in soils, fire, land use, grazing and other disturbances (Frost, 1996).

The main ecological features of miombo forests are (Chidumayo, 1997; Desanker *et al.*, 1997):

- Woodland ecosystem dominated by trees of subfamily Caesalpinioideae with predominance of *Brachystegia* and *Julbernardia* species. Woodlands are intercalated by grassy depressions and patches seasonally waterlogged;

Mocuba district forests can be broadly described as miombo forests due to the predominance of *Julbernardia globiflora* and *Brachystegia* species from the Ceasalpinioideae family (figure 11);

Figure 11 - Distribution of individuals by botanic family- Mocuba district (n=773 plots)



- Miombo areas climate is characterized by two marked rainfall seasons: the hot and wet summer where 95% of the annual rainfall occurs between November and April opposite to a dry and cooler season with almost no precipitation leaving the vegetation dry for several months of the year;

- Soils are poor in nutrients, mostly infertile limiting the agricultural potential and favoring shifting cultivation;

- Miombo as been considered a naturally fire-prone ecosystem where dry season fires are a regular and frequent event. Fire is probably the dominant disturbance factor over large areas of miombo forests since all fires kill seedlings and young regeneration. The late dry season fires are considered more intense and destructive than fires occurring in the early dry season. Despite of all awareness campaigns and civil society efforts to reduce uncontrolled forest burning it was estimated during the national forest inventory by Saket (1999) that 40% of the country's forests resources are burned annually, being the northern and central forests of the country the most affected. The same author expresses that fire and land use systems contribute to significant losses of timber trees;

- The capability of most of the miombo species to resprout from surviving stumps and rootstocks is an important survival and forest recovery strategy, that should be incorporated to forest management;

- Human populations have been occupying the miombo area for millenniums with slash and burn agriculture. In some areas, the infertile soils combined with tse-tse fly and lack of infra-structures results in very low population density.

Timber harvesting technique applied in Mozambique is selective and defined by three parameters: (1) species - only some marketable species are harvested, (2) stem quality that is, only the best trees of those commercial species are extracted, and (3) minimum harvestable diameter defined by forest regulation which requires that only adult trees can be harvested.

When the ecological characteristics are viewed within the timber extraction and industry supply perspective and harvesting method the following aspects are highlighted:

- Miombo forests are characterized by low marketable volumes per hectare that can be extracted (table 3).

Table 3 - Average commercial volume for the main commercial species –Zambézia province

Species name (common name)	Commercial volume by specie (m^3 ha^{-1})
Afzelia quanzensis (chanfuta)	0.15
Millettia stuhlmannii (jambirre)	0.47
Pterocarpus angolensis (umbila)	1.22
Swartzia madagascariensis (pau-ferro)	0.22
Dalbergia melanoxylon (pau-preto)	0.02
Pericopsis angolensis (muanga)	0.93
Khaya nyasica (umbaua)	0.09
Burkea africana (mucarala)	0.87
Combretum imberbe (mondzo)	0.06
Total	4.02

Source: Cuambe *et al.,* 2005

- Fire and shifting cultivation affect the structure of forests, tree architecture and stem quality, and consequently contribute to reduce the availability of commercial timber. As consequence, concession logging areas to supply timber industries are vast (from 20 000 to 100 000 hectares) requiring an updated forest inventory and management plan. The integration of ground data and remote sensing material is inevitable and *k*-NN algorithm may show as well a potential application at local level.

4.2.1 Reference material

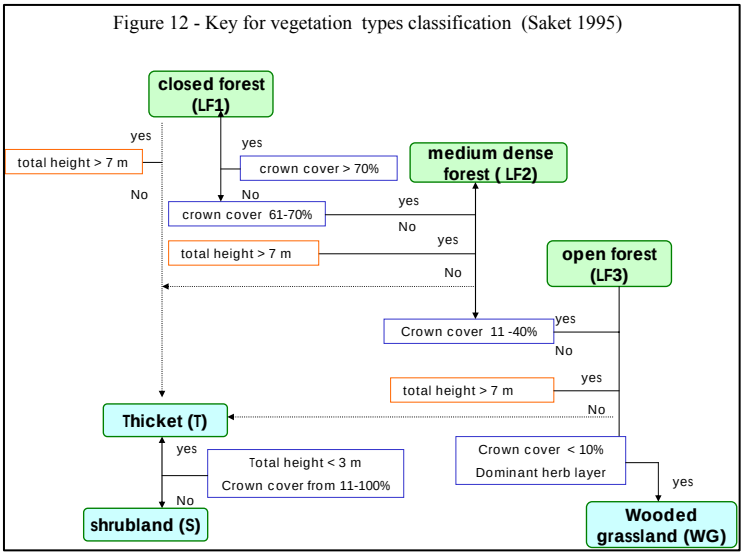
The following section describes the main results from the forest inventory work of Zambézia province and is based on the report by Cuambe *et al.* (2005) from the Forest Inventory Unit. From the 4 million hectares inventoried in 2001-2002 in Zambézia province by the Forest Inventory Unit, 499 100 ha correspond to Mocuba district (table 4).

Table 4 – Area inventoried by forest type – Mocuba district

Forest type (code)	area (ha)	%
Agriculture (A)	1 900	0.4
Grasslands (G)	1 200	0.2
Lowland dense forests (LF1)	6 100	1.2
Lowland medium dense forests (LF2)	169 200	33.9
Open Lowland Forests (LF3)	275 500	55.2
Thickets (T)	40 000	8.0
Wooded grasslands (WG)	5 200	1.0
Total	**499 100**	**100**

Forest resources were stratified by the national forest inventory unit in order to help to obtain the most efficient and representative sampling layout of the inventory clusters. The resulting land cover classes should be easily detected on the satellite images as well as comparable to the on-going country's forest classification structure adopted by Saket (1995), where forest classes are based on structural characteristics (figure 12).

Figure 12 - Key for vegetation types classification (Saket 1995)



The two stand structural characteristics (crown cover and total height) were the basis for forest types classification and countrywise vegetation maps were initially developed using this classification (1:250 000).

The forest classes defined for Zambézia were firstly created based on visual interpretation of image colours and textures according to the ability of the technicians of the forest inventory unit to interpret tones, colours, contrast, etc. Visual interpretation was complemented with NDVI value to enhance differences between dense forests and thickets, as well as with inventory pilot plots established in Derre forests. This first image interpretation resulted in a set of 19 "spectral classes".

To calculate the forest area by each major forest type and agriculture Landsat ortho-mosaic image ETM band 7, 4, and 2 was used to obtain the separation of forest lands and non-forest areas, based on unsupervised classification that produced a total of 25 spectral classes.

This allowed the first separation of forest areas and non forest areas using the satellite information and data from field checking and auxiliary maps. Non forest areas were identified, delineated and excluded (Cuambe *et al*., 2005).

The next step consisted of using the *k*-NN procedure to extrapolate the spectral value of the pixels covered by the field samples to non-sampled pixels to obtain the overall forest area spectral class map, using a 5x5 pixels (14.25 m each pixel) window. This forest map provide the mask for the *k*-NN analysis and cover an area of 500 896 ha divided into three volume strata : (1) dense forests with total volume greater than 100 m^3 ha^{-1}; (2) medium dense forests with total volume between 70 -100 m^3 ha^{-1} and (3) open forests with total volume less than 70 m^3 ha^{-1}.

The third step comprises forest types area estimation, which was done by calculating the spectral class present in each sample plot. Since samples consist of 200 m x 20 meters and cover different spectral classes within the same sample the spectral class for each subsample (25x 20 m) was extracted and the mode used to define the corresponding final spectral class for the sample. The forest type field data was added to each sample into an additional column. With both information for each sample (forest type and spectral class) the area of each forest type was calculated by counting each pixel in the respective forest type and the equivalent area of each one (pixel size of 14.25m x 14.25m). This process allowed the estimation of forest types areas presented on above table 4.

Plot volume is calculated as the sum of individual tree volumes based on the following equations and converted to hectares with an area-hectare expansion factor:

1. Total tree volume defined by the equation:

$$vt = \frac{\pi.dbh^2}{4} * ht.f \qquad \text{[eq.16]}$$

2. Commercial tree volume define by the equation:

$$Vc = \frac{\pi.dbh^2}{4} * hc * f_c \qquad \text{[eq.17]}$$

3. Number of trees per hectare or stocking:

$$N = \sum n * a_f \qquad \text{[eq.18]}$$

Where:
dbh – diameter at breast height
ht- total tree height
hc – commercial height (usually up to the first stem bifurcation)
f- stem form factor for total volume (0.65)
fc- stem commercial form factor (0.8)
n- number of trees
af – area_hectare conversion factor (2.5 for mature trees surveyed within a 0.4 ha plot and 20 for regeneration trees surveyed within a 0.05 ha plot)

With the data provided by the Forest Inventory Unit a more detailed inventory data analysis was carried out to estimate forest parameters for different variables, forest types and strata (which were not reported on the official report tables) in order to obtain a more descriptive view of Mocuba district forest resources data that could be used as reference on the later stages of applying *k*-NN to estimate the forest inventory variables.

Table 5 shows the overall forest parameters of Mocuba district within a 95% probability interval based on 167 clusters database and a cluster random sampling design where the total variance (S^2_x) is the sum of variance within (S^2_w) and between clusters (S^2_b). The variance of the mean ($S^{2}_{\bar{x}}$) is given by the equation:

$$S^{2}_{\bar{x}} = \left(\frac{N-n}{N}\right)\frac{S^2_b}{n} + \frac{S^2_w}{n.m} \qquad \text{[eq.19]}$$

Where: N: total number of primary units in the population (clusters)
n: number of primary units sampled (clusters)
M: total number of secondary units
m: number of secondary units sampled in each primary units (4 plots)
x: variable

The standard error of the mean ($S_{\bar{x}}$) is given by the square root of the variance of the mean through the equation:

$$S_x = \sqrt{\left(\frac{N-n}{N}\right)\frac{S^2_b}{n} + \frac{S^2_w}{n.m}} \qquad \text{[eq.20]}$$

The sampling error relative (E_r) and absolute (E_a)for a probability level is given by the equations:

$$Ea = \pm ts_{\bar{x}} \qquad \text{[eq.21]}$$

$$Er = \frac{\pm ts_{\bar{x}}}{\bar{x}}.100 \qquad \text{[eq.22]}$$

Where *t*: student distribution.

Table 5 - Estimated forest variables for all forest area –Mocuba district

Forest Inventory Variable	Unit	Number of clusters (N)	Average value	Relative sampling error (%)	Confidence intervals (95%		Total estimation for Mocuba forest area [1)]
					Lower level	Higher level	
Total volume	m^3 ha^{-1}	167	**99.5**	6.5	93.0	106	49 839 198 m^3
Commercial volume	m^3 ha^{-1}	167	**37.5**	7.3	34.7	40.2	18 783 617 m^3
Stocking	nr.of trees /ha	167	**321.4**	7.3	297.8	344.9	160 988 123 trees
Basal area	m^2/ha	167	**10.8**	5.4	10.2	11.3	
DBH	cm	167	**26.5**	2.1	25.9	27.1	
Ht	meters	167	**13.2**	3.1	12.8	13.6	
Hc	meters	167	**4**	5.0	3.8	4.2	

[1)] Mocuba forest area: 500 896 ha

Estimations obtained presented smaller values than the ones reported by the forest inventory unit, probably due to the exclusion of permanent sample plots from the database. Although when the whole forest area is considered it is expected to obtain a higher sampling error than for stratified forest. The error obtained for the whole forest area (non-stratified) was within the acceptable limits, that is less than 10% for all forest inventory variables (varied from 2.1% for average dbh estimation to 7.3% for commercial volume and stocking estimation).

Due to subjectivity of vegetation types classification (LF1, LF2, LF3, WG, G, T and agriculture) and heterogeneous vegetation types within the same cluster (41% of clusters with mixed vegetation types) it is not possible to use the cluster design statistical analysis for separated vegetation types, with exception of vegetation type LF2, LF3 and WG which presented more than 1 cluster, since in such sampling design, cluster (or primary units) represents the statistical unit of analysis (table 6). The

majority of clusters (70 primary units) comprised samples within different forest types (mixed vegetation type clusters).

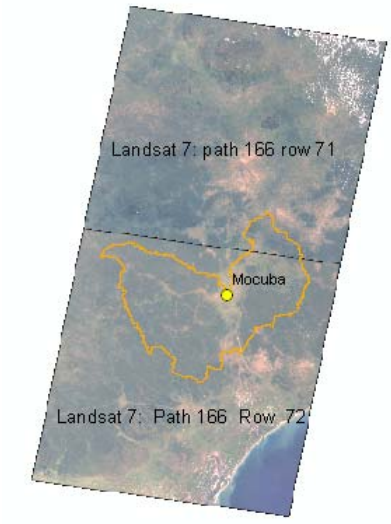
Table 6 - Number of clusters by vegetation type for Mocuba district

Vegetation types	Number of clusters (N)
LF1 – dense forests	1
Lf2 – medium dense forests	27
Lf3 – open forests	48
WG – wooded grasslands	4
G – T – A - Mixed vegetation type clusters	70
Clusters without information of vegetation type	17
Total	167

4.3 Satellite data

Two individual scenes of Landsat Enhanced Thematic Mapper are needed to cover the whole Mocuba district (Figure 13). Landsat scenes were supplied by the Forest Inventory Unit, which acquired them from Earth Satellite Corporation, a commercial provider of satellite imagery with contract with NASA.

Figure 13 - Landsat scenes covering Mocuba district



Both images are from 27 of April 2000, orthorectified, that is, corrected for the terrain displacements and satellite viewing variations with a root mean square error (RMSE) of less than 50 m (table 7).

Table 7 - Image metadata

	image metadata	
Iimage file	**p166r072_7x20000427**	**p166r071_7x20000427**
Spacecraft	Landsat7	Landsat7
Sensor	ETM+	ETM+
Acquisition date	27/april/2000	27/april/2000
WRS_path	166	166
WRS_row	72	71
Datum/ spheroid	WGS84	WGS84
Grid cell origin	Center	center
Grid increment unit	Meters	meters
Pixel size (panchromatic band)	14.250	14.250
Pixel size thermal band)	57.0	57.0
Pixel size (reflexive bands)	28.5	28.5
Orientation	North Up	North Up
Resampling interpolation method	nearest neighbour	nearest neighbour
projection	UTM	UTM
Zone	37	37
Sun azimuth	46,0978137	46,9958618
Sun elevation	46,337121	47,4864327
Output format	Geotiff	Geotiff
Pre-processing	Orthorectified	Orthorectified
Method	First order polynomial	First order polynomial
RMSE	Less than 50 m	Less than 50 m

Image bands were extracted and projected to UTM/WGS84, zone 37 south , datum -WGS84 to match the field data resulting on the following bands used (table 8):

Table 8 – Image bands used on the study

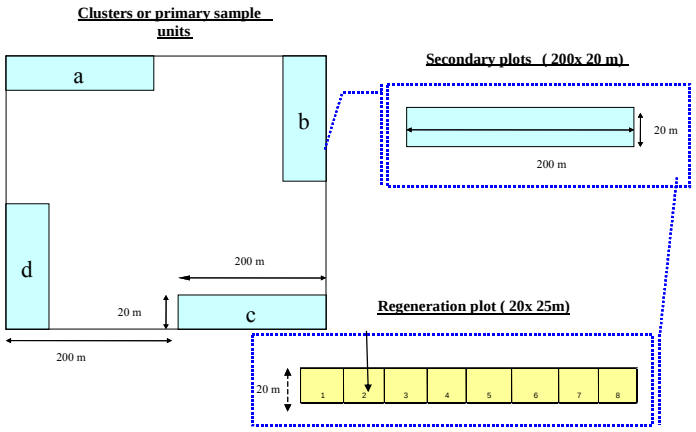
image layer	band name	spectrum	pixel size (m)	minimum value	maximum value
1	blue/green	0.45-0.52 µm	28.5	1	255
2	green	0.52-0.60 µm	28.5	29	254
3	red	0.63-0.69 µm	28.5	19	255
4	NIR	0.76-0.90 µm	28.5	1	167
5	MIR	1.55-1.75 µm	28.5	1	251

4.4 Field data

Field data used as ground truth in the *k*-NN and image analysis consisted of provincial forest inventory plots data that was collected by the national forest inventory unit from June to November 2001 within the provincial strategic forest inventory project framework. The provincial forest inventory was carried out in order to obtain the knowledge of the existing forest resources at broad provincial level and as an helping instrument for decision making on forest management strategies. Key and reliable information like total and commercial exploitable timber volumes were a primary objective of the forest inventory work.

The forest inventory design applied the cluster configuration considering that for vast and difficult to access forest areas the costs of plot location in the field are higher than plot measurement . Once the centre cluster coordinates were located in the field, four plots were measured within each cluster to reduce the location costs and the overall inventory costs (figure 14).

Figure 14 – Cluster configuration



Each cluster, also denominated *primary sample unit* comprises four plots *or secondary units* of 200 meters by 20 meters (4000 m^2). The secondary units are also known as adult trees plots because all trees with breast height diameter equal or greater than 20.1 cm are measured for: (1) total tree height (*ht*) – from the bottom to the top of the tree crown; (2) timber commercial height (*hc*)- from the tree bottom to the first brunch deviation and (3) diameter at breast height (*dbh*) – at 1,30 meters height. These plots are further sub-divided into 8 subplots of 20 x 25 meters in order to facilitate the location of *terciary plots* or established regeneration plots. In every second subplot all trees with diameter between 5 cm and 20 cm at breast height were measured for: (1) total tree height and (2) *dbh.*

This design suffered a modification from 2003 onwards where smaller plots started to be measured with 100 m length by 20 meter wide. The variables measured and sample location system remained the same. All Mocuba sample plots have 200 meters by 20 meters wide covering an area of 4000 m^2.

Clusters centre were randomly and proportionally located in each forest type in the field according to a previous forest stratification. Forest stratification was based on the visual interpretation of Landsat images spectral classes expressed by the normalized difference vegetation index - NDVI and influenced by existing Saket (1995) vegetation map and forest type classification. Saket (1995) vegetation classes were based on structural (i.e. physiognomic characteristics) definitions of the vegetation. 198 clusters (1.6 ha/each) were measured in Mocuba district producing a sampling intensity of 0.06% with an overall sampling error less than 10% achieved.

Although homogenous stratified areas were assumed to locate field sample plots, the transition between forests and non-forests classes is not always defined by a unique spectral colour and no "clean borders" are always present. This is particularly true when considering a large sample plot (200 x 20m) and a cluster of 4 samples (0.4 ha). There are cases that within the same cluster some plots are located within family agricultural fields although the other plots belong to a forest area, rendering a high volume variance within the clusters. On the other side, within a 200 meters long plot, the vegetation presents high variability among the subplots (20 x 25 m) expressed by the variation coefficient in each plot. Table 9 summarizes subplot total volume variation coefficient within each plot showing an average coefficient of variation for total volume by plot of 77% .

Table 9 – Variation coefficient by plot for total volume.

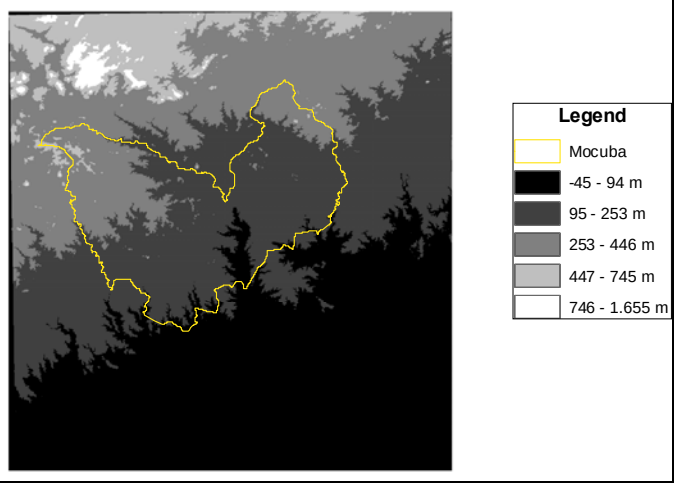
Vegetation type	Avg CV (%)	Number of plots (n)
A -agriculture	104	10
G -grassland	70	4
LF1 –dense forest	60	6
LF2 –medium dense forest	65	130
LF3 –open forest	70	263
S -shrubland	200	2
T -thicket	120	21
Wg –wooded grassland	173	20
Total	77	456

4.5 Ancillary data

The study used the following ancillary data:

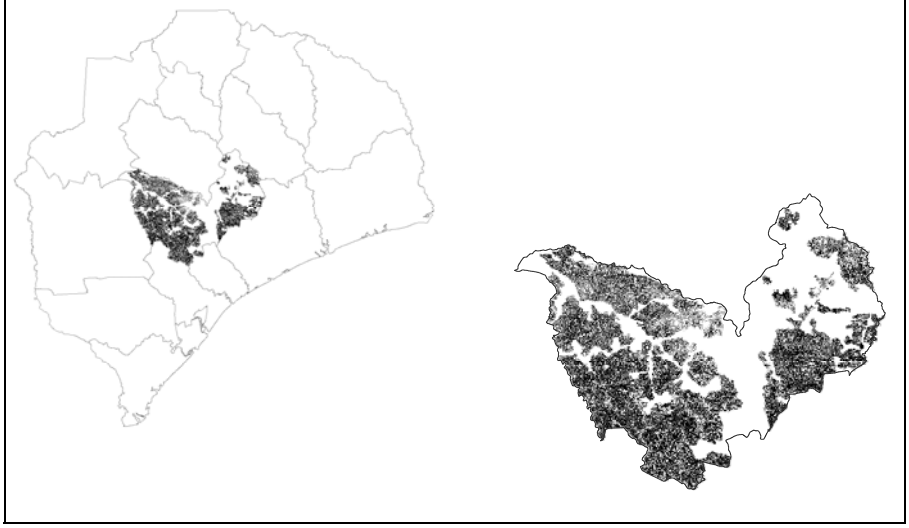
- Digitized elevation model (DEM) for Mocuba district downloaded from the GLCF (global land cover facility) site (http://glcf.umiacs.umd.edu) where elevation is given on a rectangular grid with 3 arcsec by 3 arcsec resolution (90 m) provided by the Shuttle Radar Topography Mission (SRTM) and used as auxiliary band for *k*-NN analysis (figure15).

Figure 15 - Digital earth model



- Forest map of Mocuba distritct provided by the forest inventory unit covering an area of 500 896 hectares (56.2 % of total district area) which was used as mask for *k*-NN mapping of forest variables (figure 16).

Figure 16 – Forest mask – Mocuba district.



Forest mask comprises 483 659 hectares in image 166072 and 17 236 ha in image 166071 (3.4%).

4.6 Software used

The software used in this study was designed by the Forestlab.net University Consortium in C++ programming language and operates in DOS environment. The programme is under constant improvement to fit different users needs and has the following main operations:

1) The leaveone-out procedure to extract the best *k*-NN parameters (distance and *k*) were the RMSE is estimated through cross validation within the training dataset.

2) The *k*-NN algorithm calculation for forest variables prediction and mapping, where for each query pixel the program calculates:

- Euclidean (or Mahalanobis or Mahalanobis modified distance) to all pixels that contain field data (training plots);
- Sort the training plots by Euclidean distances (or other distance selected);
- Take the *k* (*k*=1, 2 ,3 ….n) nearest training plots and estimates the desired forest variable (in this study, total, commercial volume or stocking) as weighted mean of the *k* -nearest observations;
- Proceed to the next pixel up to the completion.

Since for each query pixel the programs runs the search over all training dataset, the number of training points influences the speed of the operations. For 356 training plots the program needs about 3,5 hours to estimate one variable within the masked area.

The programme requirements regarding data used are:

- A geo-referenced image in RST format. All bands should be with data in real and binary format. Mask image must be in byte binary format for the programme to operate;
- Field plots data in text format;

4.7 Methodological steps of the study

The study methodology comprised 5 main steps: (1) data preparation, (2) forest attributes and image correlation, (3) selection and tuning of *k*-NN parameters, (4) mapping of forest attributes and (5) validation of the estimations. Table 10 shows the main steps on the methodology used to estimate forest variables with *k*-NN.

Table 10 – Methodological steps of the study

Steps	Objective	Data used	Method
1. Data preparation	Prepare/scanning and calculate data required. Band projection	Satellite image and field plots.	Spectral statistics within field plots calculation.
2. Forest attributes and image correlation.	Band selection.	▪ Plots within the image 166072. ▪ Plots within Mocuba district, with and without buffer area.	Pearson correlation coefficient. Stepwise regression analysis.
3. *K*-NN parameters selection.	Determine the best value of *k* and distance.	Training data : Pixel level - subplot 2 (20x25m) Plot level - plot (20x200m) with buffer	Leaveone-out (RMSE)
4. Mapping of forest parameters at pixel and plot level	Mapping and estimation of total volume, commercial volume and stocking	Training data . Mocuba forest mask. Dem district. Image band.	K nearest neighbour image production.
5. Validation and results comparison	Determine the bias and applicability of *k*-NN at district level	Validation data. Data from forest inventory report.	RMSE and Bias.

4.7.1 Data preparation and scanning

Mocuba district field data was scanned at plot level to withdraw from the dataset those plots with anomalies. Out of the 792 total plots available, 23 plots (2.9%) were cancelled from the initial dataset resulting into 769 plots available for Mocuba district (table 11). The vast majority of the plots (92%) are located within the image 166072 (Landsat path:166; row:72).

Table 11 - Data mining at plot level

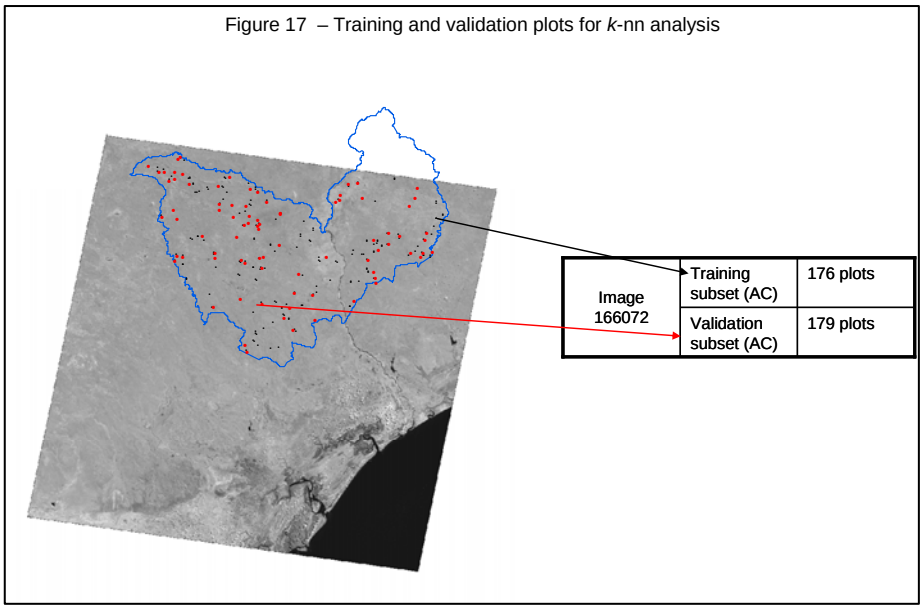
Eliminated plots	Plot number identification	Number of plots
Permanent sample plots	271; 381; 494; 533; 603; 744; 963; 1264; 1313; 2084; 2233; 3724	12
No data	2271	1
Located inside a river	3834; 2181	2
Located on a incelberg (stone)	3641; 3642; 3643; 3644	4
Located outside district	1321;1322;1323;1324	4
Total - all dataset		**769**
Total - image 166072		**711**
Total - image 166071		**58**

Due to reduced availability of Mocuba plots within image 166071, district volumes estimation and *k*-NN analysis was performed using the image 166072 since contains most of the data available for the district.

To build the *k*-NN model in order to estimate forest parameters of Mocuba district, the field plots were randomly divided into two datasets: (a) training dataset used to define the *k*-NN parameters

(distance, *k*, maximal distance, etc) through the crossvalidation procedure (leaveone out) and (b) the validation set used to compare the bias and RMSE for forest attributes estimation. *K*-NN analysis and volume estimation was done using a subset data of plots *a* and *c* (horizontal plots) that were randomly divided into 176 training plots to build the k-nn model and 179 plots to validated it (figure 17).

Figure 17 – Training and validation plots for *k*-nn analysis



4.7.2 Mapping and estimation of forest variables approaches: pixel and plot level.

Forest variables estimates were performed at two levels: (1) pixel and (2) plot level.

The pixel level approach considers the combination of satellite image data and ground data from sample plots to estimate the selected forest variables for each pixel of the image. The second sub-plot of the training dataset was selected to obtain the ground data for pixel level estimation. Second subplot was chosen because it contains the information of regeneration subplot (trees from 5 to 20 cm dbh) so that subplot data matches the overall plotdata.

Forest variables (total, commercial volume and stocking) calculated for each subplot, covering an area of 20 x 25 m, that is less than a pixel, provided the training data set. For each image pixel to be estimated and mapped, the training dataset is searched in order to identify the k nearest neighbours pixels through distance determination. Once found the k nearest neighbours, the query pixel assumes the average value of its nearest neighbours. On this approach the pixel is the central unit for analysis.

The plot level approach considers the plot size and volume to expand the average forest information contained within the plot unit to a grid of plots. In this approach, the image bands were preliminary transformed into a 7 x 1 pixel neighbouring grid (covering 5685.7 m^2) with a dynamic averaging of neighbouring pixel values in order to create a similar grid to the existing plots. *K*-NN expands the neighbouring training plot average volume/stocking to the query pixel/plot created grid. In this approach the plot is the central unit for variables estimation.

Since this procedure only allows for horizontal or vertical grid the *k*-NN mapping was performed using the horizontal training plots (plots *a* and *c* or 1 and 3), that is 176 training plots. From this same training plots the second subplot data was extracted and 173 training plots at pixel level were obtained (figure 18).

Figure 18 - Training plots for pixel and plot level k-nn estimation

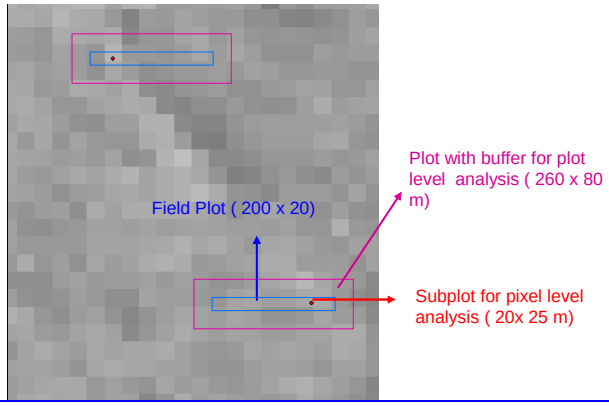


Table 12 shows the statistics summary of training data used in this study, where there is less variation in forest variables for plot units (200 x 20m) than for subplots, that are smaller than a size of a Landsat ETM pixel (28.5x28.5).

Table 12 - Statistics summary of training data

Variables statistics	*Pixel level training data n=173 plots)*			*Plot level training data (n=176 plots)*		
	Stocking ($N\ ha^{-1}$)	*VT* ($m^3\ ha^{-1}$)	*Vc* ($m^3\ ha^{-1}$)	*Stocking* ($N\ ha^{-1}$)	*VT* ($m^3\ ha^{-1}$)	*Vc* ($m^3\ ha^{-1}$)
Average	272.9	97.1	37.9	307.5	97.7	35.7
Median	260	77.7	28.1	266.3	97.9	33.6
Moda	0	0	0	430.0	77.6	0
Standard deviation	187.7	99.3	38.9	193.5	56.6	23.4
Variance	35259.3	9877.7	1518.0	37424.8	3208.9	546.9
Range	880	759.5	242.6	937.5	274.6	113.0
Minimum	0	0	0	0	0	0
Maximum	880	759.5	242.6	937.5	274.6	113.0

4.7.3 Validation of *k*-NN estimates

After mapping the forest variables through the *k*-NN algorithm a map of continuous variables (volumes, number of trees) is obtained. Estimation results were checked by comparing the estimated pixel volume or plot volume with the actual values obtained in the field using the validation dataset. The reliability of the estimation was measured by means of standard error (RMSE) and bias (equations 9 and 11).

K-NN forest variables estimates are than compared to the reference material, that is, estimates obtained by the traditional forest inventory for Mocuba district.

5. Results

5.1 *K*-NN parameters calibration

5.1.1 Correlation between band spectral information and field plots variables.

The Pearson's correlation coefficient is a single number that describes the degree of relationship between two variables. Table 13 illustrates the correlation between the various bands and dendrometric parameters for all plots within the image 166072. Results show that total volume is strongly correlated to basal area and moderate correlated to total height since both constitute main variables for total volume calculations. Total volume and commercial volume are strongly correlated (r=0.85). Forest commercial volume is strongly correlated to total volume, basal area and commercial height. Stocking or forest density expressed in number of trees per hectare and other forest variables showed weak correlation with dendrometric variables (total and commercial volume) except for diameter at breast height which showed a moderated correlation with stand density.

Correlation analysis among band spectral information and field plots variables showed a weak correlation in all cases. Green band was the band that shows the highest correlation for stand basal area, commercial and total volume (r= -0.37). For dbh, total and commercial height both red and green band shows the highest correlation coefficient (between 0.33 and 0.34). The correlation matrix portrayed in table 13 contains as well, useful information about the redundancy of bands. It is observed that in the visible spectrum there is strong correlation among bands: green band shows the strongest correlation with red (r=0.92), blue (r= 0.80) and middle infrared band (r=0.80). The middle infrared band shows correlation with the red band. Near infra red band shows almost no correlation with all other bands (table 13). In practice it makes more sense to select the bands with minimum redundancy, that is, no correlation among them, to allow for better visualization and separation of features.

Table 13 – Pearson correlation coefficient between forest variables and mean digital number for plots without buffer (n=834) for all landsat image 166072

								Mean digital number by band					
Variables	*Vt*	*Vc*	*Stocking*	*BA*	*DBH*	*Ht*	*Hc*	*blue*	*Green*	*red*	*NIR*	*MIR*	*NDVI*
total volume	1												
Comercial volume	0.85	1											
Stocking	0.31	0.18	1										
Basal area	0.86	0.77	0.59	1									
DBH	0.43	0.45	-0.48	0.24	1								
Total height	0.69	0.58	-0.22	0.33	0.65	1							
Comercial Height	0.49	0.76	-0.25	0.31	0.59	0.68	1						
MEAN DN_blue	-0.24	-0.33	-0.06	-0.27	-0.27	-0.21	-0.34	1					
MEAN DN_green	-0.37	-0.38	-0.11	-0.38	-0.33	-0.32	-0.33	0.80	1				
MEAN DN_red	-0.35	-0.36	-0.07	-0.33	-0.33	-0.34	-0.33	0.77	0.92	1			
MEAN DN_NIR	-0.004	-0.02	-0.10	-0.10	0.02	0.04	-0.03	0.03	0.20	0.08	1		
MEAN DN_MIR	-0.24	-0.20	-0.07	-0.25	-0.24	-0.22	-0.18	0.66	0.80	0.85	0.25	1	
MEAN NDVI	0.22	0.25	0.05	0.18	0.14	0.20	0.22	-0.40	-0.39	-0.44	0.27	-0.30	1

Summing up, the correlation between spectral band information and forest variables was weak when the all image was analysed, due to heterogeneous conditions and coverage of vast area.

The same analysed was performed **only** for Mocuba district plots, considering the spectral information contained within the plots of 200 meters x 20 meters (**no buffer area around plot**). Table 14 summarizes Pearson correlation coefficient results showing a similar behaviour but a slight

improvement on the correlation between the spectral information (mean digital number) and forest variables, where the total volume presents a better correlation with the green and red band (r=0.47).

Table 14 - Pearson correlation coefficient between forest variables and mean digital number for plots without buffer for Mocuba district (n=711)

								Mean digital number by band					
	Vt	*Vc*	*Stocking*	*AB*	*DBH*	*HT*	*HC*	*Blue*	*Green*	*Red*	*NIR*	*MIR*	*NDVI*
Total volume	1												
Commercial volume	0.88	1											
Stocking	0.53	0.39	1										
Area basal	0.91	0.83	0.73	1									
DBH	0.43	0.44	-0.15	0.34	1								
HT	0.65	0.56	0.04	0.43	0.76	1							
HC	0.49	0.72	-0.06	0.38	0.66	0.71	1						
MEAN_blue	-0.36	-0.34	-0.23	-0.39	-0.35	-0.29	-0.31	1					
MEAN_green	-0.47	-0.43	-0.27	-0.49	-0.42	-0.41	-0.39	0.88	1				
MEAN_red	-0.47	-0.43	-0.26	-0.47	-0.42	-0.42	-0.41	0.88	0.95	1			
MEAN_NIR	-0.10	-0.11	-0.12	-0.17	-0.11	-0.08	-0.12	0.16	0.38	0.22	1		
MEAN_MIR	-0.36	-0.32	-0.21	-0.38	-0.33	-0.31	-0.30	0.84	0.87	0.91	0.31	1	
MEAN_NDVI	0.43	0.39	0.20	0.40	0.37	0.39	0.35	-0.77	-0.74	-0.86	0.26	-0.72	1

A similar analysis was performed considering the median digital number and the majority of DN (mode) within the plots but no improvement was found.

In order to account for inevitable plot positional displacements and image registration errors, the same analysis was performed considering the spectral information within plots of 260 meters x 80 meters (**30 meters buffer area around the plot**) and considering only the **Mocuba district plots**. Table 15 shows that no improvement was found if a buffer area of 30 meters is considered.

Table 15 -Pearson correlation coefficient between forest variables and mean digital number for plots with 30 meters buffer area for Mocuba district (n=711)

								Mean digital number by band					
	Vt	*Vc*	*Stocking*	*AB*	*DBH*	*HT*	*HC*	*Blue*	*Green*	*Red*	*NIR*	*MIR*	*NDVI*
Total volume	1												
Commercial volume	0.88	1											
Stocking	0.53	0.39	1										
Area basal	0.91	0.83	0.73	1									
DBH	0.43	0.44	-0.15	0.34	1								
HT	0.65	0.56	0.04	0.43	0.76	1							
HC	0.49	0.72	-0.06	0.38	0.66	0.71	1						
MEAN_blue	-0.35	-0.34	-0.23	-0.38	-0.33	-0.29	-0.31	1					
MEAN_green	-0.45	-0.42	-0.27	-0.47	-0.39	-0.40	-0.39	0.91	1				
MEAN_red	-0.45	-0.43	-0.26	-0.46	-0.40	-0.41	-0.41	0.91	0.96	1			
MEAN_NIR	-0.13	-0.14	-0.13	-0.20	-0.10	-0.09	-0.13	0.21	0.40	0.27	1		
MEAN_MIR	-0.36	-0.33	-0.22	-0.38	-0.31	-0.31	-0.30	0.88	0.88	0.93	0.34	1	
MEAN_NDVI	0.45	0.40	0.22	0.43	0.40	0.41	0.36	-0.67	-0.66	-0.71	0.08	-0.61	1

One reason might be that on a such a large plots the mean digital number values remain mostly the same with or without buffer area. In practice terms, there are always displacement errors and therefore spectral data information derived from plots with buffer area was used for *k*-NN model building at plot level although not much correlation improvement was achieved.

On the other hand when subplot of 20x25 meters (less then a pixel) data is analysed the correlation between field data and image spectral information reduces due to heterogeneity of subplot information and subplot locational errors influence at pixel level. Table 16 shows that the correlation coefficient for total volume at its best is -0.26 for the red band.

Table 16 -Pearson correlation coefficient between forest variables and digital number for training subplots (25x20m)

				Central point digital number					
	Vt	*Vc*	*stocking*	*blue*	*green*	*red*	*nir*	*mir*	*ndvi*
Total volume	1.00								
Commercial vol.	0.91	1.00							
stocking	0.49	0.40	1.00						
Blue	-0.15	-0.15	-0.19	1.00					
Green	-0.19	-0.14	-0.19	0.09	1.00				
Red	-0.26	-0.28	-0.22	0.52	0.10	1.00			
Nir	-0.08	-0.06	-0.05	0.14	0.14	0.23	1.00		
Mir	-0.23	-0.23	-0.25	0.41	0.06	0.51	0.22	1.00	
Ndvi	0.22	0.25	0.19	-0.45	-0.03	-0.89	0.25	-0.41	1.00

5.1.2 Image band selection through multiple regression

In order to select the most appropriate bands to estimate forest variables a stepwise multiple regression analysis was performed for three main forest variables: (i) total volume. (2) commercial volume and (3) stocking, since they are the most important timber related variables.

The coefficient of determination, symbolized by r^2, is the square of correlation coefficient and provides a measure of the strength of correlation, that is, give us the percentage of variance (fluctuation) of one variable that is predictable from another variable. and was used as the main criterion to judge the linear regression models.

For total volume estimation

Multiple Regression Analysis shows that the R-Squared statistic is low, that is indicates that the model as fitted explains 28.01 of the variability in total volume. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 26.7 (table 17).

Table 17 – Regression analysis summary for total volume as dependent variable

Dependent variable: total Volume (m^3 ha^{-1})

Parameter	Standard Estimate	Error	T Statistic	P-Value
CONSTANT	-390.88	205.873	-1.89864	0.0584
MEAN DN _blue	15.1396	4.63052	3.26953	0.0012
MEAN DN_green	-9.95234	4.9061	-2.02856	0.0433
MEAN DN_red	-2.41601	4.03741	-0.59840	0.5500
MEAN DN_nir	-1.6081	1.23389	-1.30327	0.1933
MEAN DN_mir	1.24902	1.2279	1.0172	0.3098
MEAN DN_NDVI	471.515	97.6208	4.83007	0.0000

```
Analysis of Variance
 ------------------------------------------------------------------------------
 Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
 ------------------------------------------------------------------------------
Model                 315414.0      6      52569.0      22.64       0.0000
Residual              810537.0    349      2322.46
 ------------------------------------------------------------------------------
     Total (Corr.)          1.12595E6    355

R-squared = 28.0131 percent
R-squared (adjusted for d.f.) = 26.7755 percent
Standard Error of Est. = 48.1919
Mean absolute error = 37.0407
Durbin-Watson statistic = 1.60665 (P=0.0001)
Lag 1 residual autocorrelation = 0.195193
```

The model suggested with the stepwise analysis with a criterion to enter and remove variables of 3.5 (F to enter = 3.5 and F to remove =3.5) with the three main significant bands (blue, green and NDVI) is:

*total Volume = -540.375 + 18.8543*MEAN DN_blue - 13.6656*MEAN DN_green + 428.534*MEAN DN_NDVI*

R-squared = 27.58 percent
R-squared (adjusted for d.f.) = 26.96 percent

For commercial volume estimation

Regression analysis showed for commercial volume a lower adjusted r-squared than for total volume (table 18).

Table 18 –Regression analysis summary for commercial volume as dependent variable

```
------------------------------------------------------------------------------
              Dependent variable: Commercial volume  (m³ ha⁻¹)

------------------------------------------------------------------------------
                           Standard          T
Parameter              Estimate         Error       Statistic        P-Value
------------------------------------------------------------------------------
CONSTANT               -66.5501        88.1761      -0.754741         0.4509
MEANblue                4.15384        1.98326        2.09445         0.0369
MEANg                   -2.0976         2.1013       -0.99824         0.3189
MEANred                -2.95237        1.72923       -1.70733         0.0887
MEANnir                -0.92848        0.52848       -1.75689         0.0798
MEANmir                 1.14193       0.525915        2.17133         0.0306
MEANNDVI                131.732        41.8112        3.15065         0.0018
------------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                  43631.3      6      7271.88      17.07       0.0000
Residual               148687.0    349      426.038
------------------------------------------------------------------------------
    Total (Corr.)          192319.0    355

R-squared = 22.687 percent
R-squared (adjusted for degrees of freedom.) = 21.3578 percent
Standard Error of Estimation = 20.6407
Mean absolute error = 15.7067
Durbin-Watson statistic = 1.72108 (P=0.0042)
Lag 1 residual autocorrelation = 0.138193
```

When a model with statistically significant bands was constructed using the same “F-to remove” and “ F-to-enter” criteria. the suggested final equation is:

*Commercial volume = 105.237 - 3.42317*MEANred - 1.42139*MEANnir + 1.5696*MEANmir + 129.648*MEANndvi*

```
R-squared = 21.71 percent
R-squared(adjusted for d.f.) = 20.82 percent
```

For forest density estimation (stocking)

Forest density, expressed as number of trees per hectare, is an important forest variable due to its simplicity for field mensuration and as a forest management variable that can be easily altered and manipulated by human activity in the forests, show almost no correlation with mean digital number of image bands (table 19).

Table 19 – Regression analysis summary for stocking as dependent variable

Dependent variable: STOCKING (Number of trees per hectare)

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1044.06	838.222	1.24556	0.2138
MEANblue	-9.96343	18.8533	-0.52847	0.5975
MEANg	11.1083	19.9754	0.556097	0.5785
MEANred	-19.1251	16.4385	-1.16344	0.2454
MEANnir	-10.2132	5.02385	-2.03294	0.0428
MEANmir	6.55063	4.99947	1.31027	0.1910
MEANNDVI	809.517	397.467	2.03669	0.0424

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1.40315E6	6	233858.0	6.07	0.0000
Residual	1.34367E7	349	38500.4		
Total (Corr.)	1.48398E7	355			

R-squared = 9.45529 percent
R-squared (adjusted for d.f.) = 7.89865 percent
Standard Error of Est. = 196.215
Mean absolute error = 156.039
Durbin-Watson statistic = 1.43725 (P=0.0000)
Lag 1 residual autocorrelation = 0.281311

Those variables with a P-value greater or equal to 0.10, that is, almost all variables in the full model (blue band, green, red and MIR) are not statistically significant at 90% and higher confidence (table18). The suggested model was constructed only with NIR and NDVI bands but still no big improvement was achieved (r squared adjusted = 8.27%)

*STOCKING = 555.293 - 9.49389*MEANnir + 1161.71*MEANNDVI*

R-squared = 8.79 percent
R-squared (adjusted for d.f.) = 8.27 percent

As a conclusion, the following bands were the most appropriate for estimation of the three main forest variables (table 20):

Table 20 - Summary of band selection

Forest variable	Bands suggested	Coefficient of determination (r^2 adjusted - %)
Total volume	Blue Green NDVI	27.5
Commercial volume	Red NIR MIR NDVI	20.82
Stocking	NIR NDVI	8.27

Multiple linear regression models confirmed the relative small relation between forest variables and the spectral information on the image through low coefficient of determination. Therefore, the suggested bands indicated on the above table 19 can serve only as an indication since, no linear relationship seems to exist between forest variables and spectral information in each band. The *k*-NN alghorithm which do not require any explicit model and relationship between dependent/independent variables offer an alternative option for forest variables estimation on this case.

.

5.1.3 Image band selection through *k*-NN and crossvalidation

k-NN software allows the crossvalidation by predicting the selected forest variable in one omitted sample at a time and calculates the *k*-NN prediction error when it predicts the omitted sample, expressed by root mean square error. The best band combination is the one that provides the lowest root mean square error.

Table 21 summarizes the several alternatives of bands and auxiliary information considering a fixed number of neighbours of 10 as reference for comparison of various options.

Table 21 - Band selection alternatives based on k-NN crossvalidation

Alternatives tested	Input data for *k*-NN local tuning			Cross validation statistics (reference k =10)
	Bands	Auxiliary band	Spectral data used as input file	
1.	B,G,R, NIR, MIR	DEM	central plot point	r=0.072; rmse=58.89
2.	B,G,R, NIR, MIR	DEM	Training Data _ plots with buffer	r=0.379; rmse=53.00
3.	B,G,R,NIR,MIR,DEM	-	Training Data _ plots with buffer	r=0.313; rmse=54.50
4.	B,G,R,NIR,MIR,NDVI	DEM	Training Data _ plots with buffer	r=0.229; rmse=56.80
5.	NIR, R, G	DEM	Training Data _ plots with buffer	r=0.270; rmse=56.82
6.	B, G, NDVI	DEM	Training Data _ plots with buffer	r=0.392; rmse=53.28

In the first and second alternative the only variation is the type of spectral information used for *k*-NN crossvalidation. On the first alternative, it is used as input file the spectral information calculated by the *k*-NN software which considers only the digital number of the central plot point. On the second alternative, the input file considers the information contained in all pixel and its average within a plot with a buffer area around it. Results shows a considerable improvement if it is used the information contained in all plot area and buffer instead of the central point digital number.

The third alternative considers the use of the digital earth model (DEM) as a band, but no improvement was achieved when compared to alternative 2. The model performed better if DEM is used as auxiliary band.

In the alternative 4, the use of NDVI information as an input band did not show any improvement.

The fifth alternative tested the use of the false color combination (432=NIR,R,G) but no improvement was obtained. When the bands suggested by the linear model were used (sixth alternative which includes blue, green and NDVI) the best r was obtained confirming the highest correlation between total volume and those bands. But, root mean square error did show improvement when compared with the possibility of using all bands (alternative 2) Therefore, based on the criteria of obtaining the smallest RMSE the alternative 2 was selected as the best option, that is, use of all 5 bands (B, G, R, NIR and MIR) is recommended with the digital earth model (DEM) as auxiliary band and data from al pixels contained within the plot with buffer as the input datafile.

5.1.4 Selecting the distance

The *k*-NN software allows three types of distance selection and testing: (1) Euclidean distance also known as the minimum distance between two points. (2) Mahalanobis distance which considers the covariance of each band and (3) Mahalanobis modified which gives different weights to the bands. with inverse weight, that is, bands with high variability receive less weight.

Either for pixel or plot level the Euclidean distance resulted in the smallest RMSE if a reference k=10 was compared (figure 21). Pixel estimations showed a higher RMSE (slightly more than double) when compared to plot level estimation for 10 neighbours value.

Figure 21 – Best distance - total volume estimation

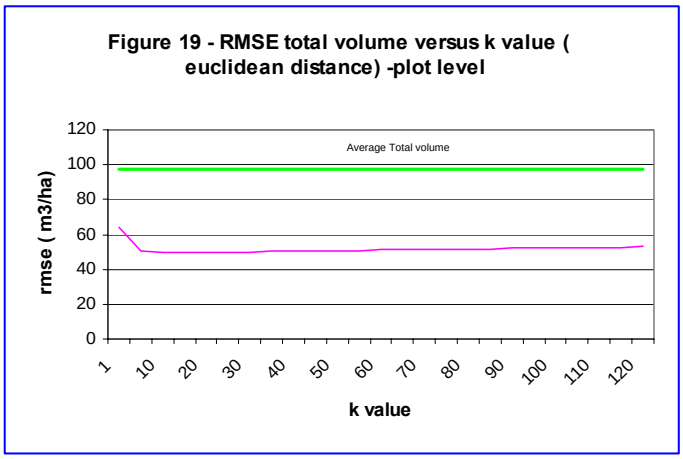
Pixel level total volume estimation			
k	RMSE-mahalanobis modified	RMSE-Mahalanobis	RMSE-euclidean
1	131,54	137,35	134,82
2	118,27	116,54	116,18
3	109,72	109,75	109,61
4	106,77	106,17	105,32
5	105,90	105,76	102,55
6	105,15	104,78	103,09
7	103,38	103,90	101,81
8	102,45	102,77	101,83
9	102,14	102,79	101,45
10	101,30	101,67	100,70
11	101,37	101,35	100,20
12	101,23	101,46	99,88
13	100,22	100,39	99,18
14	98,84	98,99	98,93
15	98,63	98,76	98,52
16	98,82	98,68	98,40
17	98,92	98,55	99,17
18	98,60	98,71	98,93
19	98,52	98,78	98,72
20	98,92	98,82	98,05

Plot level total volume estimation			
K	RMSE mahalanobis_ modified	RMSE mahalanobis	RMSE minimum distance
1	66,28	67,52	63,80
2	58,06	59,11	55,53
3	51,86	55,50	52,45
4	53,00	54,31	51,78
5	52,15	53,26	50,78
6	51,18	51,77	49,53
7	50,59	51,80	50,17
8	50,69	51,23	49,72
9	50,53	51,00	49,27
10	50,44	50,83	49,40
11	50,55	50,50	49,14
12	50,25	50,59	49,11
13	50,21	50,65	49,49
14	50,20	50,75	49,75
15	50,24	50,56	49,91
16	50,31	50,85	49,83
17	50,62	50,87	49,92
18	50,91	51,14	50,03
19	50,96	51,49	49,97
20	51,10	51,41	49,91

The same behaviour was detected for commercial volume and stocking estimation and therefore the Euclidean distance was used for *k*-NN forest variables prediction.

5.1.5 Selecting the number of neighbors

The value of k, that is the number of neighbours to be included in the calculation of predictions, is one of the local calibration procedures to increase the accuracy of estimations. The number of neighbours was determined for both *k*-NN approaches: pixel level and plot level and in both cases a typical k value versus RMSE curve referred on the literature was obtained with large initial decreases of RMSE and a more gradual decrease in RMSE as the optimal k approaches. Then, a very gradual increase in RMSE towards the overall mean as *k* approaches the number of observations (figure 19).



The criteria used to determine the best number of neighbours was based on the relative RMSE value smaller than 0.5% obtained from cross validation - since the leaveone out error estimator is considered nearly unbiased estimator of the true error – between *k* and *k*-1 sustained for a interval ranging for *k* and *k*+5 (Katila, 2004).

Considering the criteria above mentioned, for total volume estimation the *k* value at pixel level stabilizes at much higher number of neighbours **-24 -** than for plot level **-15.** Both values were used for *k*-NN prediction since are the first *k* value from where the rate of RMSE decrease is less than 0.5% and sustained for an interval of 5 k (table 22).

Table 22 – Selection of best number of neighbours (k) for total volume estimation

k	Pixel level		Plot level	
	RMSE (m^3/ha) - euclidean	Δ RMSE%	RMSE (m^3/ha) - euclidean	Δ RMSE%
1	134.82		63.80	
2	116.18	-13.8%	55.53	-12.95%
3	109.61	-5.66%	52.45	-5.56%
4	105.32	-3.91%	51.78	-1.27%
5	102.55	-2.63%	50.78	-1.94%
6	103.09	0.53%	49.53	-2.45%
7	101.81	-1.24%	50.17	1.28%
8	101.83	0.02%	49.72	-0.90%
9	101.45	-0.37%	49.27	-0.91%
10	100.70	-0.74%	49.40	0.26%
11	100.20	-0.50%	49.14	-0.51%
12	99.88	-0.32%	49.11	-0.06%
13	99.18	-0.70%	49.49	0.77%
14	98.93	-0.25%	49.75	0.53%
15	98.52	-0.41%	**49.91**	0.32%
16	98.40	-0.12%	49.83	-0.17%
17	99.17	0.78%	49.92	0.18%
18	98.93	-0.24%	50.03	0.23%
19	98.72	-0.21%	49.97	-0.13%
20	98.05	-0.68%	49.91	-0.11%
21	97.93	-0.12%		
22	97.42	-0.52%		
23	98.04	0.64%		
24	**98.03**	-0.01%		
25	97.72	-0.32%		
26	97.56	-0.16%		
27	97.35	-0.22%		
28	97.52	0.17%		

For commercial volume estimation the same general behaviour was found: for pixel level the RMSE reduction stabilizes at higher *k* values (**18**) than for plot level (**k=11**) and these values were used for k-NN commercial volume prediction and mapping (table 23).

Table 23 – Selection of best number of neighbours (k) for commercial volume estimation

	Pixel level		Plot total	
k	RMSE (m^3 ha^{-1}) - euclidean	Δ RMSE%	RMSE (m^3 ha^{-1}) - euclidean	Δ RMSE%
1	51.00		29.18	
2	46.23	-9.4%	25.05	-14.12%
3	43.27	-6.40%	23.62	-5.72%
4	42.52	-1.73%	23.26	-1.52%
5	42.24	-0.66%	22.74	-2.23%
6	41.69	-1.30%	22.19	-2.45%
7	41.02	-1.61%	22.08	-0.46%
8	40.83	-0.46%	21.99	-0.43%
9	40.66	-0.42%	21.78	-0.94%
10	40.12	-1.33%	21.53	-1.17%
11	39.90	-0.55%	**21.45**	-0.36%
12	39.77	-0.33%	21.46	0.04%
13	39.55	-0.55%	21.46	-0.01%
14	39.26	-0.73%	21.44	-0.06%
15	39.36	0.25%	21.47	0.13%
16	39.25	-0.28%	21.33	-0.64%
17	39.00	-0.64%	21.30	-0.13%
18	**39.03**	0.08%	21.37	0.32%
19	39.02	-0.03%	21.31	-0.27%
20	38.94	-0.21%	21.33	0.07%
21	39.00	0.15%		
22	38.96	-0.10%		
23	38.98	0.05%		
24	39.04	0.15%		
25	39.00	-0.10%		
26	39.09	0.23%		
27	39.10	0.03%		
28	39.02	-0.20%		
29	38.94	-0.21%		
30	38.93	-0.03%		

Table 24 shows that for stocking estimation the number of neighbours is similar either for pixel or plot level being equal to 11 and 13, respectively.

Table 24 – Selection of best number of neighbours (k) for stocking estimation

K	Pixel level		Plot total	
	RMSE (m^3 ha^{-1}) - euclidean	Δ RMSE%	RMSE (m^3 ha^{-1}) - euclidean	Δ RMSE%
1	271.12		240.40	
2	230.23	-15.1%	182.98	-24%
3	213.7	-7.18%	182.90	-0.05%
4	209.31	-2.05%	195.50	6.89%
5	204.84	-2.14%	196.32	0.42%
6	202.94	-0.93%	190.54	-2.95%
7	200.56	-1.17%	188.47	-1.08%
8	197.91	-1.32%	188.32	-0.08%
9	197.17	-0.37%	187.61	-0.38%
10	196.07	-0.56%	187.55	-0.03%
11	**195.13**	-0.48%	186.88	-0.36%
12	195.79	0.34%	185.17	-0.92%
13	195	-0.40%	**185.56**	**0.21%**
14	195.02	0.01%	184.89	**-0.36%**
15	195.22	0.10%	185.36	**0.25%**
16	194.91	-0.16%	185.06	**-0.16%**
17	193.64	-0.65%	184.19	**-0.47%**
18	193.27	-0.19%	184.06	-0.07%
19	193.44	0.09%	183.64	-0.23%
20	192.66	-0.40%	182.98	-0.35%

5.2 Forest stratification

It is expected that if plots are aggregated within homogeneous vegetation characteristics the correlations between variables will improve. Correlation analysis was performed for training data at plot level considering:

Stratification by vegetation types formed two main strata: (i) mixed forests and agriculture and (ii) forests. Mixed forests and agriculture combine agricultural fields, grasslands, shrublands, woodlands and thickets where field classification and low total volumes constitutes the common criteria of this stratum. The forest stratum comprises open, medium dense and dense forests. Table 25 shows the main statistics of training plot data where only 156 plots presented information regarding vegetation type.

Table 25 – Summary statistics by vegetation type and strata

Stratum	Vegetation types	By vegetation type				By strata			
		number of plots (n)	Average Stocking (n/ha)	Average Vt (m^3/ha)	Average Vc (m^3/ha)	number of plots (n)	Average Stocking (n/ha)	Average Vt (m^3/ha)	Average Vc (m^3/ha)
Mixed forests and agriculture	A- agriculture	5	34	18.0	4.4	21	85	16.2	4.5
	G – grasslands	2	30	7.5	1.85				
	S – shrubland	1	45	5.6	0.7				
	WG- wooded grassland	7	144	20.4	6.1				
	T- thicket	6	84	14.6	4.1				
Forests	LF1 – closed forests	1	352	85.6	42.3	135	337	111.7	41.1
	LF2 – medium dense forests	51	356	121.2	48.4				
	LF3- open forests	83	325	106.2	36.6				
Total		156				156			

Most of the training plots that were classified as mixed forests (agriculture. grasslands. shrublands and others) are located outside the forest mask utilized.

With the Pearson correlation matrix illustrated on table 26 it is possible to observe that stratification increased the correlation between bands and forest attributes of mixed forests (plots with low volume) while correlation values within the strata " forests" which included only the forests showed no improvement when compared to no stratified vegetation.

Table 26 - Pearson correlation coefficient for stratification by vegetation type : forests and mixed forests strata

training data - Plot level						no stratification						
	Vt	***Vc***	***N***	***DBH***	***HT***	***HC***	***AB***	***blue***	***green***	***red***	***nir***	***mir***	***NDVI***
Vt	1.00												
Vc	0.87	1.00											
N	0.49	0.36	1.00										
DBH	0.43	0.47	-0.27	1.00									
HT	0.70	0.56	0.03	0.66	1.00								
HC	0.51	0.74	-0.09	0.63	0.65	1.00							
AB	0.89	0.83	0.70	0.32	0.43	0.40	1.00						
Blue	-0.41	-0.41	-0.25	-0.45	-0.40	-0.44	-0.45	1.00					
Green	-0.22	-0.21	-0.22	-0.16	-0.21	-0.19	-0.30	0.15	1.00				
Red	-0.48	-0.42	-0.30	-0.48	-0.52	-0.45	-0.50	0.82	0.25	1.00			
Nir	-0.25	-0.24	-0.30	-0.08	-0.17	-0.21	-0.33	0.24	0.67	0.30	1.00		
Mir	-0.48	-0.45	-0.31	-0.41	-0.47	-0.46	-0.49	0.71	0.27	0.80	0.41	1.00	
NDVI	0.01	-0.06	0.00	0.13	0.09	-0.01	0.00	-0.33	0.16	-0.39	0.25	-0.19	1.00
training data - Plot level						Strata: forests							
Vt	1.00												
Vc	0.81	1.00											
N	0.31	0.13	1.00										
DBH	0.37	0.45	-0.66	1.00									
HT	0.65	0.46	-0.24	0.48	1.00								
HC	0.34	0.70	-0.38	0.55	0.49	1.00							
AB	0.84	0.76	0.58	0.15	0.20	0.17	1.00						
Blue	-0.16	-0.20	0.02	-0.28	-0.09	-0.19	-0.14	1.00					
Green	0.06	0.03	-0.04	0.12	0.13	0.09	-0.02	-0.20	1.00				
Red	-0.26	-0.20	-0.03	-0.24	-0.25	-0.19	-0.20	0.64	0.00	1.00			
Nir	-0.08	-0.08	-0.21	0.11	0.05	-0.04	-0.18	-0.07	0.61	0.06	1.00		
Mir	-0.32	-0.32	-0.17	-0.13	-0.20	-0.26	-0.30	0.51	0.11	0.72	0.29	1.00	
NDVI	0.19	0.13	-0.08	0.27	0.24	0.15	0.08	-0.60	0.30	-0.86	0.45	-0.50	1.00
training data - Plot level					Strata: mixed forests								
Vt	1.00												
Vc	0.97	1.00											
N	0.49	0.40	1.00										
DBH	0.42	0.50	-0.12	1.00									
HT	0.61	0.64	0.01	0.83	1.00								
HC	0.46	0.58	-0.19	0.85	0.84	1.00							
AB	0.93	0.87	0.73	0.33	0.46	0.31	1.00						
Blue	-0.41	-0.50	-0.23	-0.41	-0.43	-0.56	-0.42	1.00					
Green	-0.43	-0.50	-0.02	-0.40	-0.51	-0.57	-0.34	0.13	1.00				
Red	-0.48	-0.53	-0.20	-0.50	-0.52	-0.55	-0.48	0.91	0.10	1.00			
Nir	-0.45	-0.54	-0.17	-0.27	-0.32	-0.48	-0.37	0.57	0.67	0.39	1.00		
Mir	-0.50	-0.59	-0.10	-0.58	-0.60	-0.69	-0.42	0.85	0.22	0.83	0.52	1.00	
NDVI	0.38	0.40	0.17	0.46	0.46	0.45	0.40	-0.80	0.12	-0.94	-0.08	-0.73	1.00

Although not much improvement was found from the separation of forests and mixed forests in training data correlation . the *k*-NN prediction was applied only to the masked area. which covered mostly the forest strata, leaving aside mixed forests, agricultural lands and non forests areas.

5.3 Forest variables mapping and estimation

5.3.1 Total volume mapping and estimation

Total volume mapping and estimation was made only restricted to the forest mask provide by the FIU and no further stratification within mask. *K*-NN parameters tuned on the first part of this study were used for map production and estimations based on pixel and plot approach.

Table 27 shows the training data statistics used showing that for pixel level. total volume ranged from 0 to 759.5 m^3 ha^{-1} while for plot level the maximum plot volume was 274.6 m^3 ha^{-1}. *k* value for pixel level total volume prediction was much higher (*k*=24) than for plot level (*k*=15). When higher *k*-values are used the *k*-NN averaging produces an underestimation of higher values and overestimation of smaller volumes and the full variance of data is disturbed. In both approaches average total volume was overestimated when compared to mean district total volume in 9% and 4%, respectively for pixel and plot approach.

Table 27 – Training data and k-NN prediction for total volume (m^3 ha^{-1})

k-NN prediction approach	**Training data used**				**k value**	**Estimated parameters with *k*-NN**			**Reference data (forest inventory)**
	n	$\bar{x}$	*x min*	*x max*		$\bar{x}$	*x min*	*x max*	$\bar{x}$
pixel	173	97.1	0	759.5	24	108.5	0	525.1	99.5
plot	176	97.7	0	274.6	15	103.7	0	206.2	

Total volume mapping is presented in annex 1 and 2.

5.3.2 Commercial volume mapping and estimation

Commercial volume presented a smaller variation than the total volume but due to the averaging of *k*-NN the estimated maximum values were also smaller than the maximum value found in training data. Mean *k*-NN estimates are very similar to the one obtained for the district area (table 28).

Table 28 –Training data and k-NN prediction for commercial volume (m^3 ha^{-1})

k-NN prediction approach	**Training data used**				**k value**	**Estimated parameters with *k*-NN**			**Reference data (forest inventory)**
	n	$\bar{x}$	*x min*	*x max*		$\bar{x}$	*x min*	*x max*	$\bar{x}$
pixel	173	37.9	0	242.6	18	36.3	0	140.9	37.5
plot	176	35.7	0	113.0	11	37.3	0	89.6	

Commercial volume mapping is presented in annex 3 and 4.

5.3.3 Stocking mapping and estimation

Forest density expressed as number of trees per hectare was predicted using *k* equal to 11 and 13 for pixel and plot level approach. While the pixel prediction was underestimated the plot level estimation was nearly the same as the forest inventory (table 29).

Table 29 – Training data and k-NN prediction for stocking (N ha^{-1})

k-NN prediction approach	**Training data used**				**K value**	**Estimated parameters with *k*-nn**			**Reference data (forest inventory)**
	n	$\bar{x}$	*x min*	*x max*		$\bar{x}$	*x min*	*x max*	$\bar{x}$
pixel	173	272.9	0	880	11	241.1	0	626.3	321.4
plot	176	307.5	0	937.5	13	326.4	0	659.9	

Stocking map is presented in annex 5 and 6

5.4 Validation

Validation plots within mask area provided the database for calculating bias and RMSE using the central pixel information and plots/subplot volumes measured in the field.

The RMSE is a widely used criteria to evaluate the estimations given by the *k* nearest neighbors method (Sironen *et a*l., 2001). Absolute root mean square error estimated by the cross validation (leaveone-out method) during the *k*-NN parameters tuning was compared to the one obtained by validation dataset. Table 30 shows that cross validation gives a fairly good indication of the expected absolute mean error since values are approximately similar to those obtained with the validation dataset.

Table 30 – Absolute RMSE obtained with cross validation and validation

Estimation level	Variables	units	Cross validation Training set	Validation data set
			RMSE	RMSE
Pixel	Total volume	m^3 ha^{-1}	98.0	82.9
Plot			49.9	49.4
Pixel	Commercial volume	m^3 ha^{-1}	39.0	35.9
Plot			21.4	23.7
Pixel	Stocking	N ha^{-1}	195.1	202.3
Plot			185.6	199.6

The relative RMSE allows the comparison of different approaches and variables. In all three forest attributes relative errors for pixel predictions were clearly higher than those of the plot estimations. Table 31 shows that for pixel estimations relative error varied from 76 to 99% while for plot estimations the relative error reduced from 47 to 63%. Of the three variables analyzed, total volume predictions presented the lowest relative RMSE followed by stocking and commercial volume .

Table 31 – Errors of the forest variables prediction for pixel and plot level estimates

Variables	units	n	estimation	Bias	relative bias	t bias		RMSE	Relative RMSE
total volume	m^3 ha^{-1}	137	pixel	13.4	12.4%	1.91	not sig.	82.9	76.4%
		138	plot	-1.5	-1.4%	-0.34	not sig.	49.4	47.6%
commercial volume	m^3 ha^{-1}	137	pixel	-1.6	-4.0%	-0.50	not sig.	35.9	99.0%
		138	plot	-4.2	-11.2%	-2.11	sig.	23.7	63.5%
Stocking	N ha^{-1}	137	pixel	-64.7	-26.8%	-3.95	sig.	202.3	83.9%
		138	plot	-6.3	-1.9%	-0.37	not sig.	199.6	61.1%

Bias represents the average difference between estimated and observed values. Bias was significantly different from zero for a 95% confidence level for commercial volume estimations at plot level and stocking predictions at pixel level. A not significant average bias indicates that the algorithm constitutes a good predictor of the observed average values. Bias is higher for total volume estimations of plots with high volumes due to the averaging effect of the neighbouring method (table 32).

Table 32 –Total volume bias for plot estimations by volume class

Total volume classe m^3 ha^{-1}	Avg. bias	n
0-34	68.97	10
34-67	37.08	19
67-100	17.53	35
100-134	-10.45	43
134-176	-29.82	18
167-200	-68.65	9
200-330	-151.33	4
Total average	-1.5	138

6. Conclusions and discussion

The aim of this study was to investigate whether *k*-NN method could be used to estimate and map the African miombo forest variables utilizing strategic regional forest inventories sample plots as reference material. over large area. as vast as a district area. Two main procedures were tested. pixel and plot level. to estimate and mapping three main forest variables: total volume. commercial volume and forest density.

The *k*-NN method is well studied in Nordic countries and since 1990 is operational in Finnish national forest inventory to integrate ground information and remotely sensed data. The African miombo forests composition comprises several species and forests are fragmented with patches of grasslands and shifting cultivation agricultural fields creating an heterogeneous environment. The aim of this study was to explore the possibility of using the *k*-NN method to estimate forest variables on such forests based on the already existing forest inventory field sample data.

The first step was the calibration of *k*-NN parameters to the district training data to obtain the most appropriated distance and number of neighbours for each variable to be predicted. Predictions obtained are encouraging since they are within the ranges cited in literature by many Authors. The total volume estimation using the average spectral values for a sample plot of 200 x 20 m (with 30 meters buffer around it) gave the best estimation results. Obviously averaging in such large plots reduces the effects of plots dislocations, image geometric errors and GPS accuracy, allowing better results. The opposite happens when subplots less than a pixel size are used to extrapolate the ground information, pixel by pixel to all image. Crucial conditions in this case are that the coordinates of the field plots are accurately determined and that images are correctly rectified. Pixel dislocations and geometric accuracy are maximized and as a result the correlation between image features and ground information is also smaller influencing the accuracy of pixel estimates.

Consequently, as expected and mentioned in the literature the relative RMSE's for pixel–level estimates are always rather high (Kilpeläinen and Tokola, 1999; Mäkelä and Pekkarinen, 2004; Franco-Lopez *et al.,* 2001). Expected results mentioned in the literature reveals that for pixel-level total volume estimates errors varies from 50-80%. In this study a relative RMSE of 76% is reported for single pixels total volume prediction.

The accuracy of estimations of volume is always better in an aggregated level than in a per-pixel level (Tokola, 2000). As expected, *k*-NN plot estimations showed better results than per pixel level and RMSE dropped to 47-63%.

Total volume is the variable that presents the higher correlation with image features and, therefore, showed the best results when compared to other estimated forest variables: commercial volume and forest density. Commercial volume, although highly correlated with total volume, presented either for pixel or plot level estimations the higher error among the three variables tested.

When considering average bias and mean values the appropriateness of the algorithm predictor is judged. In most cases, a not significant bias for 95% probability level was found when the reliability of estimates was tested through the validation dataset. At plot level predictions *k*-NN overestimates the total district volume in 4% and forest stoking in 1% showing that *k*-NN is a fairly good predictor algorithm. Other criteria to be considered, when the appropriateness of the algorithm is judged, is the amount of time required for processing the estimates. The running *k*-NN software time is highly dependent on the number of training plots, since for each image pixel to be estimated and mapped the training dataset must be searched in order to obtain the *k* nearest neighbours. If a training dataset of 1000 - 5000 plots is to be used the slow speed of nearest neighbour identification can make the use of *k*-NN software impractical, unless some improvements are made on the existing software and on the processing speed of computers. At fairly large district area, such as Mocuba, the programme could be runned with reasonable time speed (1.5 to 4 hours), and did not constitutes a serious constraint.

The main constraint of using *k*-NN predictions remains on the error of estimates where at the best the RMSE of the total volume estimates was 47% for plot volume estimations, still far higher than the allowable sampling error for forest inventory estimates (< 20%). Estimates accuracy can be improved through stratification to account for spatial variations and dependence of forest variables to vast areas such as district level. A maximum search radius and quota restrictions could be introduced in order to improve predictions. Leaveone-out performed at training dataset showed only a slightly improvement of the RMSE, reducing the expectations of major improvements through horizontal and vertical distance stratification for Mocuba district, which presents an almost flat topography.

Sources of errors, such as inaccurate GPS plot location were not investigated. Field data forms indicate an average GPS precision of 4 meters which was accounted for on the plot analysis with buffer area and on the pixel size (800 m^2) relation to subplot area (500 m^2). A high *k* value can control to some extent field plot dislocations but cannot reduce map errors, satellite rectification errors or sampling error.

Despite of the estimation errors obtained for *k*-NN forest variables predictions, the present study considers the obtained results encouraging as an alternative for forest inventory continuous variables mapping. The method is considered more statistically oriented than the traditional classification-based approach with satellite images (Tomppo, 1999). The visualization of forest variables is an important instrument for strategic decision making, specially if coupled with error information assessed by cross-validation/validation usually not present on the maps commonly produced. Forest maps can be easily produced with *k*-NN and certainly are more informative than circular dots information type maps. The possibility of knowing their estimation error enhances the potential as visual instruments for decision making.

Reliability of estimates obtained was within literature range despite of heterogeneous forests and inherent data collection errors and inaccuracies expected in tropical world, proofing that the method can also be used to estimate and mapping African miombo forest variables with ground data provided by strategic forest inventories. The application of k-NN method to other forest areas and country regions should always be done with prior *k*-NN parameters calibration to local data and present study parameters should not be extrapolated to other areas than Mocuba.

More studies on the refinement of the method considering the existing forest conditions and data available should be continued, specially regarding the stand-level variables estimation and horizontal / vertical stratification.

References:

Andersen, G. (1998) Classification and estimation of forest and vegetation variables in optical high resolution satellites: a review of methodologies. IIASA. Interim report 98-085. 25 p.

Ardö, J. (1992) volume quantification of coniferous forest compartments using spectral radiance recorded by Landsat Thematic Mapper. *International Journal of Remote Sensing 13*: 1779-1786

Bannary, A., Morin, D., Huete, A. R. and Bonn, F. (1995) A review of vegetation indices. *Remote sensing Reviews* 13: 95-120.

Babuska, I., Nobile, F., Oden, T.J. and Tampone, R. (2003) Reliability, uncertainty, estimates validation and verification. Transcription of presentation at the workshop on elements of predictability. J. Hopkins University. 15 p.

Boyd, D.S., Foody, G.M. and Ripple, W.J. (2002) Evaluation of approaches for forest cover estimation in the Pacific Northwest, USA, using remote sensing. *Applied Geography 22*: 375-392.

Chen, D. and Rogan, J. (2004) Remote sensing technology for mapping and monitoring land-cover and land-use change. *Progress in Planning Journal 61*: 301-325.

Campbell, J. B. (2002) Introduction to remote sensing. 3rd edition. Taylor and Francis press. 620 p.

Chidumayo, E. (1997) Miombo ecology and management: an introduction. Stockholm Environment Institute. 166 p.

Chirici, G., Corona, P., Marchetti. M., Maselli F. and Bottai L. (2003) Spatial distribution modeling of forest attributes coupling remotely sensed imagery and GIS techniques. Modelling Forest Systems. 10 p.

Cuambe, C., Norjamäki, I., Sedano, F. and Viitanen. J. (2005) Inventario florestal da provincia da Zambézia. DNFFB. PMSR. 33 p.

Curran, P.J. and Williamson, H.D. (1985) The accuracy of ground data used in remote-sensing investigations. *International Journal of Remote Sensing 6*: 1637-1651.

Desanker, P., Frost, P., Justice, C. and Scholes, R. (eds.) (1997) The miombo network: framework for terrestrial transect study of land-use and land cover change in the miombo ecosystems of central Africa –International Geosphere-Biosphere Programme / IGBP report 41. Stockholm. 109 p.

ETC Foundation (1987) Biomass assessment – A study of the SADCC region.

Fazakas, Z. N. and Olsson, M.H. (1999) Regional forest biomass and wood volume estimation using satellite data and ancillary data. *Agricultural and Forest Metereology 98-99*: 417-425.

Foody, G., Cutler, M., Mcmorrow, J., Pelz, D., Tangki, H., Boyd, D. and Douglas I. (2001) Mapping the biomass of Bornean tropical rain forest from remotely sensed data. *Global Ecology and Biogeography 10*: 379-387.

Fowler, T. (2004) Use of cross validation in forecast verification. National Center for Atmospheric Research. In http ://www .bom.gov.au /bmrc/wefor /staff/eee/verif/ Workshop2004 /presentations/1.2_Fowler.pdf

Franklin, S.E. (2001) Remote sensing for sustainable forest management. CRC Press. 407 p.

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

Frost, P. (1996) The ecology of miombo woodlands. In: *The miombo in transition : woodlands and welfare in Africa*. Campbell, B. (Ed.). Centre for International Research (CIFOR). p. 11-57.

Gutierrez-Osuna, R. (no date) Introduction to pattern recognition. Wright State University. http://research.cs.tamu.edu/prism/lectures.htm

Halme, M. and Tomppo, E. (2001) Improving the accuracy of multi-source forest inventory estimates by reducing plot location error – a multi-criteria approach. *Nordic trends in forest inventory management planning and modeling*. Proceedings of 2001 SNS meeting. 101-109.

Haapanen, R., Ek, A. Bauer, M. and Finley, A. (2004) Delineation of forest/nonforest land use classes using nearest neighbor methods. *Remote Sensing of Environment. 89:* 265-271.

Haara, A.. Maltamo, M. and Tokola, T. (1997) The k-nearest neighbour method for estimating basal-area diameter distribution. *Scandinavian Journal of Forest Research* 12: 200-208.

Horler, D.N.H. and Ahern, F. J. (1986) Forestry information content of thematic mapper data. *International Journal of remote. Sensing. 7* : 405-428.

Howard, J. (1991) Remote sensing of forest resources-theory and application. Chapman & Hall. 420 p.

Holmgren, J., Joyce, S., Nilsson, M. and Olsson, H. (2000) Estimating stem volume and basal area in forest compartments by combining satellite image data with field data. *Scandinavian Journal Forest Research. 15*: 103-111.

Holmström, H. (2001) Data acquisition for forestry planning sensing based sample plot imputation. Doctoral thesis. Swedish University of Agricultural Sciences. Umeå.. 41 p.

Huete, A. R. (1988) A Soil-Adjusted Vegetation Index (SAVI). *Remote Sensing of Environment. 25*: 295-309.

Jordan, C.F. (1969) Derivation of leaf area index from quality measurements of light on the forest floor. *Ecology 50*: 663-666.

Katila, M. (2004) Controlling the estimation errors in the Finnish multisource national forest inventories. Academic dissertation. Finnish Forest Research Institute. *Research paper 910*. 35 p.

Katila. M. and Tomppo. E. (2001) Selecting estimation parameters for Finnish multisource National Forest Inventory. *Remote Sensing of Environment Journal* 76: 16-32

Kaufman, Y. J. and Tanre, D. (1992) Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. In : Proceedings IEEE Symposium. 92. IEEE. New York. 261-270.

Kilkki, P. and Päivinen, R. (1987) Reference sample plots to combine field measurements and satellite data in forest inventory. In remote sensing-aided forest inventory. Proceedings of seminars organized by SNS. Hyytiällä. Finland. Research note 19: 209-215.

Kilpeläinen, P. and Tokola, T. (1999) Gain to be achieved from stand delineation in Landsat TM image based estimates of stand-volume. *Forest Ecology and Management 124* : 105-111.

Korhonen, K. and Kangas, A. (1997) Application of Nearest-Neighbor Regression for Generalizing Sample Tree Information. *Scandinavian Journal Forestry Research* 12 : 97-101.

Kowero, G., Campbell, B.. and Sumaila, U. (eds). (2003) Policies and governance structures in woodlands of southern Africa. *Center for International Forestry Research*. 439 p.

Kriegler, F. J., Malila, W. A., Nalepka, R. F. and Richardson, W. (1969) Preprocessing transformations and their effects on multispectral recognition. In: Proceedings of the Sixth International Symposium on Remote Sensing of Environment. University of Michigan. Ann Arbor. MI. pp. 97-131.

Lillesand, T. and Kiefer, R. W. (1979) Remote sensing and image interpretation. John Wiley & sons. Inc. 612 p.

Loetsch, F., Zohrer, F. and Haller, K.E. (1973) Forest inventory –volume I. BLV Verlagsgesellschaft. Munchen. 436 p.

Lu, D., Mausel, P., Brondízio, E. and Moran, E. (2004). Relationships between forest stand parameters and Landsat TM spectral responses in Brasilian Amazon Basin. *Forest Ecology and Management* 198: 149-167.

MADEBRAS/FUPEL/MADEMO (1982) Inventário florestal – Niassa.

Malinen, J. (2003) Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement. *Silva Fennica 37* (1): 109-120.

Mallinis, G., Koutsias, N., Apostolos, M. and Kateris, M.(2004) Forest parameters estimation in European mediterrenean landscape using remotely sensed data. *Forest Science Journal* 50 (4): 450-460.

Maltamo, M. and Kangas, A. (1998) Methods Based on K-Nearest Neighbor Regression in the Prediction of Basal Area Diameter Distribution. *Canadian. Journal Forestry Research* 28:1107-1115.

Malleux, J. (1980) Avaliação dos recursos florestais da Republica Popular de Moçambique. Projecto Moz /76/007.

Maselli, F., Bonora, L. and Battista, P. (2001) Integration of spatial analysis and fuzzy classification for estimation of forest parameters in mediterranean areas. *Remote sensing Reviews* 20: 71-88.

Mäkelä, H. and Pekkarinen, A. (2004) Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. *Forest Ecology and Management* 196:245-255.

McRoberts, R.E., Nelson, M.D.. and Wendt, D.G (2002) Stratified estimation of forest area using satellite imagery inventory data and the k-nearest neighbors technique. *Remote Sensing of Environment* 82: 457-468.

Nilsson, M. (2002) Deriving nationwide estimates of forest variables for Sweden using Landsat ETM+ and field data. Department of Forest Resources Management and Geomatics. Swedish University of Agricultural Sciences. Paper presented at ForestSAT Symposium. Edimburgh. 7 p.

Perrin, C., Michael, C. and Andreassian, V. (2001) Does a large number of parameters enhance model performance. Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hidrology 242*: 275-301.

Rajaniemi, S., Tomppo, E., Roukolainen, K. and Tuomisto, H. (2005) Estimating and mapping pteridophyte and Melastomataceae species richness in western Amazonian rainforests. *International Journal of Remote Sensing. 26* (3): 475-493.

Reese, H., Nilsson, M., Sandstrőm, P. and Olsson, H. (2002) Applications using estimates of forest parameters derived from satellite and forest inventory data. *Computers and Electronics in Agriculture 37*: 37-55.

Richardson, A. J. and Wiegand, C. L. (1977) Distinguishing vegetation from soil background information. *Photogrammetric Engineering and Remote Sensing. vol. 43*:1541-1552.

Rouse, J.W., Haas, R.H., Schell, J.A., and Deering, D.L. (1973) Monitoring vegetation systems in the great plains with ERTS. Third ERTS Symposium. NASA SP-351 I: 309-317.

Saket, M., Taquidir M. and Banze C. (1995) Methodology and results of the forestry vegetation mapping at 1:250 000. FAO/UNDP/MOZ/92/013. 22 p.

Saket, M. (1999) Tendencies of forest fires in Mozambique. Annex 4 to " proposal of a model of integrated forest management plan for timber concession of Maciambose - Cheringoma North of Sofala. Ministério de Agricultura e pescas. DNFFB/ADB-ETC.

Sengupta, N. (2003) Detection and prediction of biodiversity patterns as a rapid assessment tool in the tropical forest of East Usambara, Eastern Arc Mountains, Tanzania. Academic dissertation. 119 p.

Sironen, S., Kangas, A., Maltamo, M. And Kangas, J. (2001) Estimating individual tree growth with the k-nearest neighbour and k-most similar neighbour methods. *Silva Fennica 35*(4):453-467.

Silviconsult (1984) Inventory of the Resources in Southern Cabo Delgado – Mozambique. Direcção Nacional de Florestas e Fauna Bravia. 77 p.

Schreuder, H. T., Gregoire, T. G. and Wood, G. B. (1993) Sampling methods for multiresource forest inventory. Wiley. New York. U.S.A.

Tokola, T. (2000) The influence of field sample data location on growing stock volume estimation in Landsat TM-based forest inventory in eastern Finland. *Remote Sensing of Environment Journal 74*: 422-431.

Tomppo, E. (1999) Forest inventory – a challenge for statistics. Bulletin of the International Statistical Institute. ISI99. The 52nd Session. Proceedings. Book 1: 3-6.

Tomppo E. Korhonen. K.T.. Heikkinen J. and Yli-Kojola. H. (2001) Multi-source inventory of the forests of the Hebei forestry bureau. Heilongjiang. China. *Silva Fennica 35*(3): 309-328

Tomppo. E.. Nilsson. M.. Rosengren. M. Aalto. P. and Kenedy. P. (2002) Simultaneous use of Landsat-TM and IRS-IC WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote sensing of Environment Journal 82*: 156-171.

Tommola, M., Tynkkynen, M., Lemmetty, J., Harstela, P. and Sikanen, L. (1999) Estimating the Characteristics of a Marked Stand Using K-Nearest-Neighbor Regression. *Journal of Forest Engineering 10 (2):*75-81.

Tuominen, S. and Poso, S. (2001) Improving multi-source forest inventory by weighting auxiliary data sources. *Silva Fennica 35 (2):*203-214.

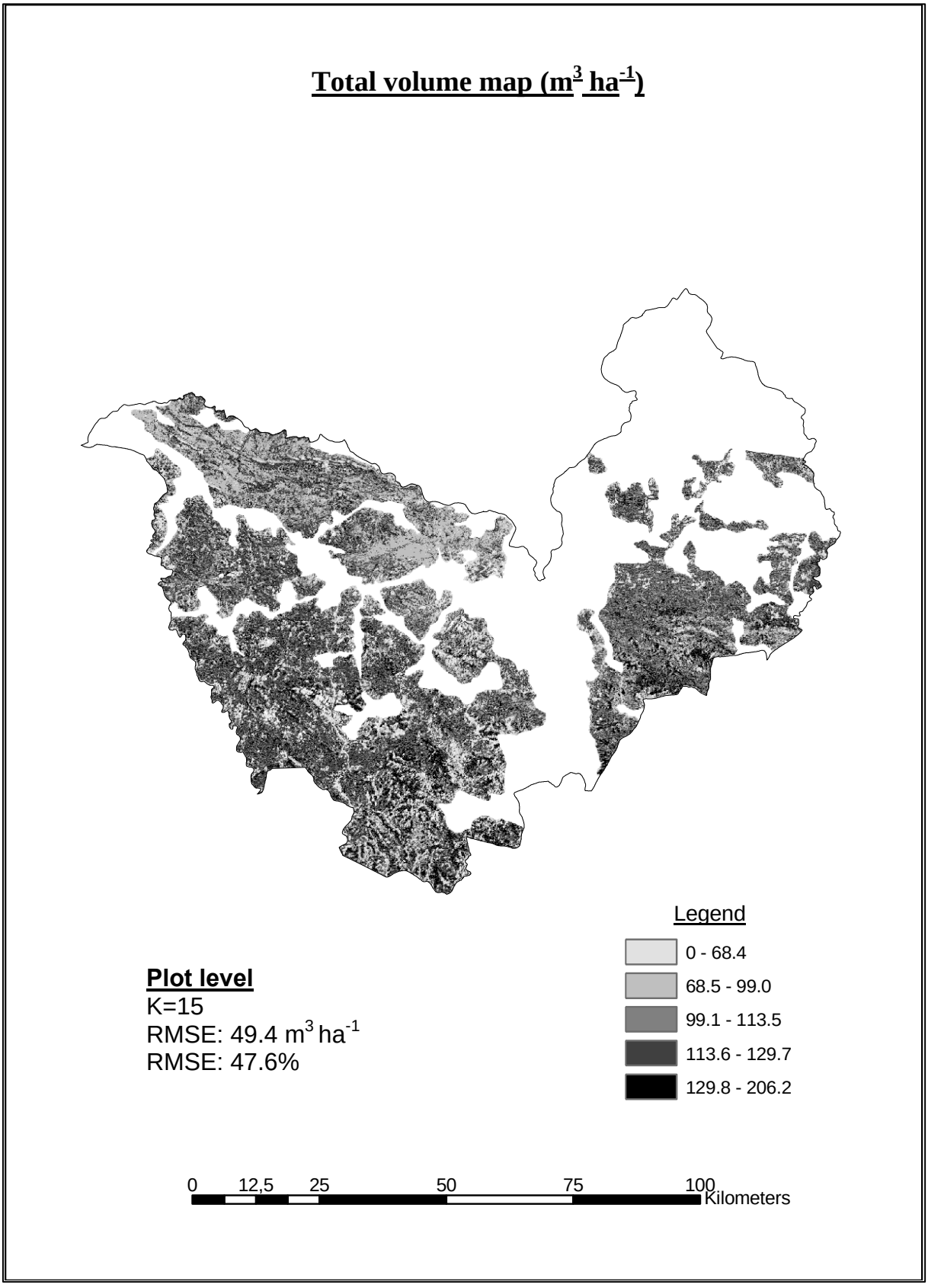
Wainwright, J. and Mulligan, M. (2004) Environmental modelling – finding simplicity in complexity. Environmental monitoring and modelling research group. Department of Geography. King's College London. John Wiley & Sons. Ltd. 408 p.

Wagacha., P.W. (2003) Instance-based learning: *K*-nearest neighbour- notes for ICS320 foundations of learning and adaptative systems. Institute of computer Science. University of Nairobi.

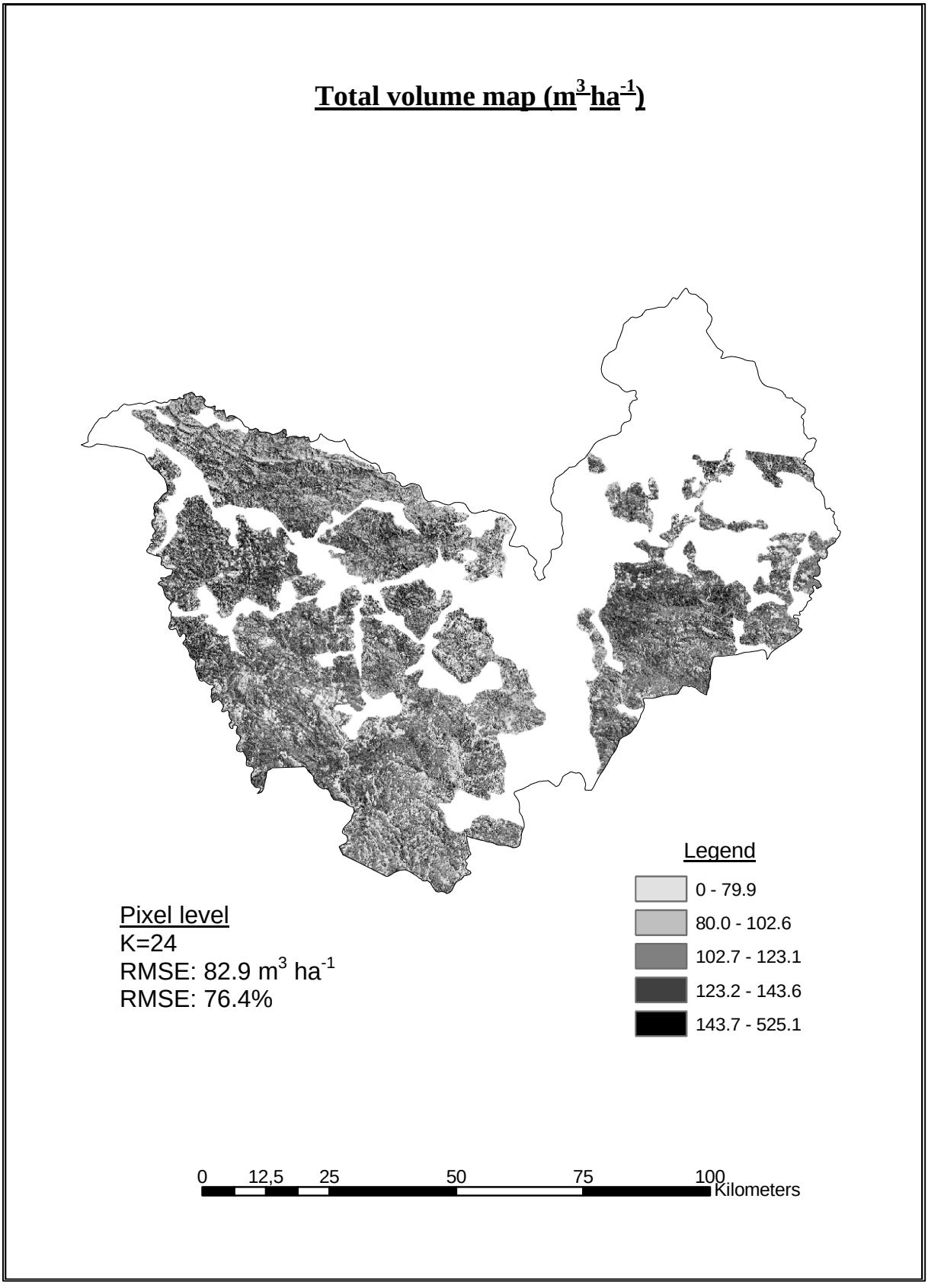
Wallerman, J. (2003) Remote sensing aided spatial prediction of forest stem volume. Doctoral thesis. Swedish University of Agricultural Sciences. Umeå.

WWF-SARPO (2001) Conserving the miombo ecoregion. Reconnaissance Summary. WWF- Southern Africa Regional Programme Office. Harare. Zimbabwe. 24p.

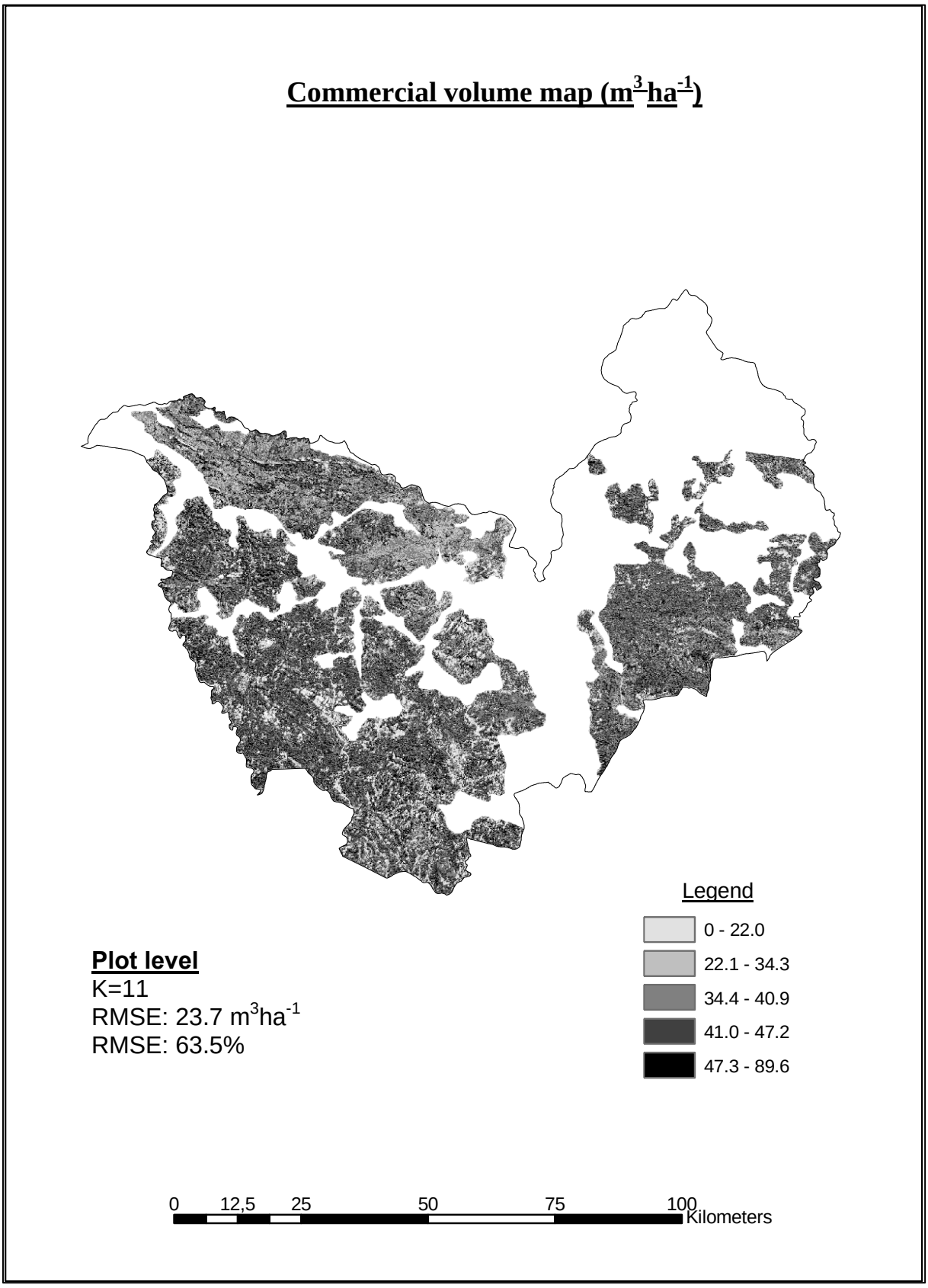
ANNEX 1 – Total volume mapping (plot level)



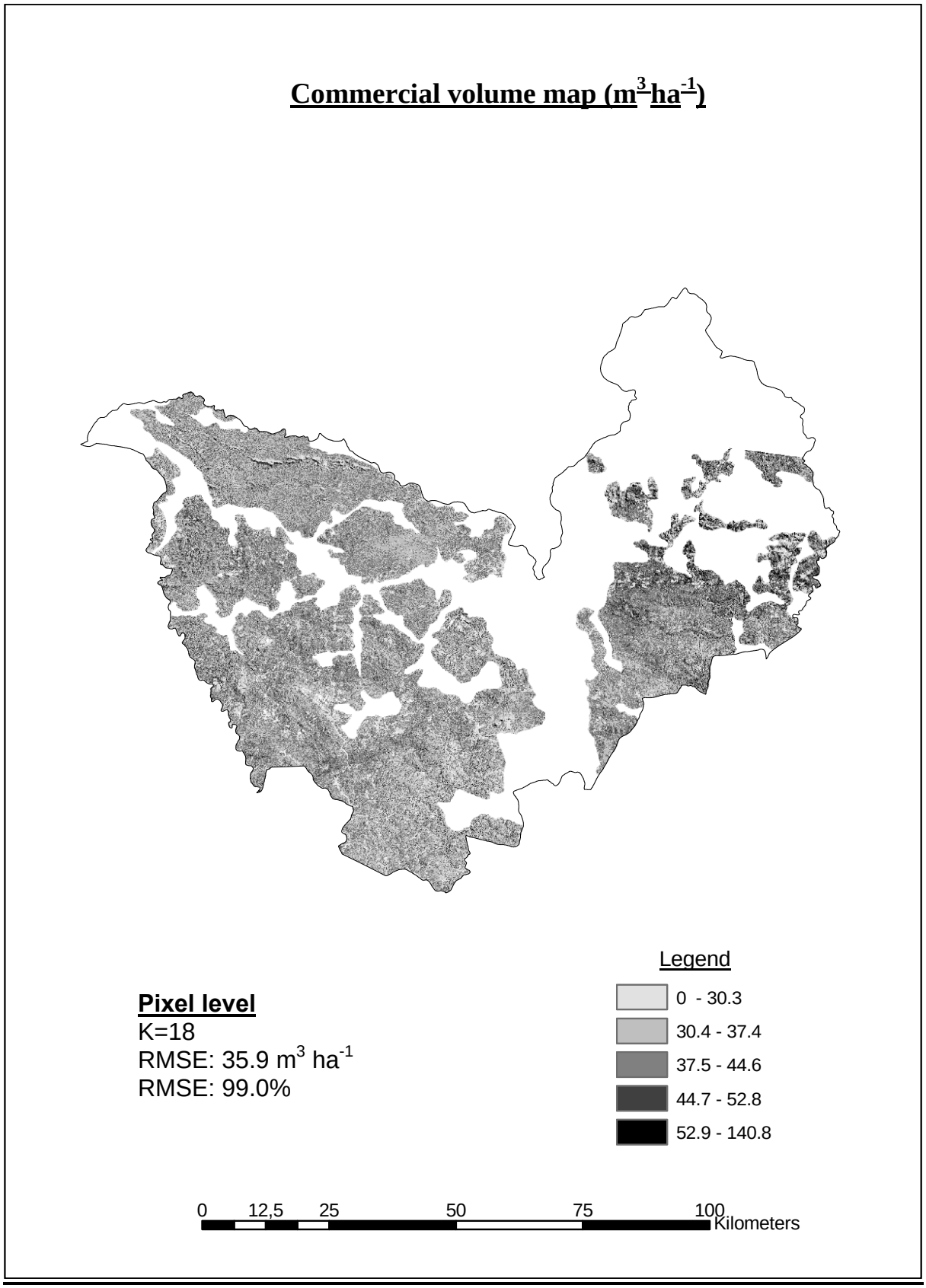
ANNEX 2 – Total volume mapping (pixel level)



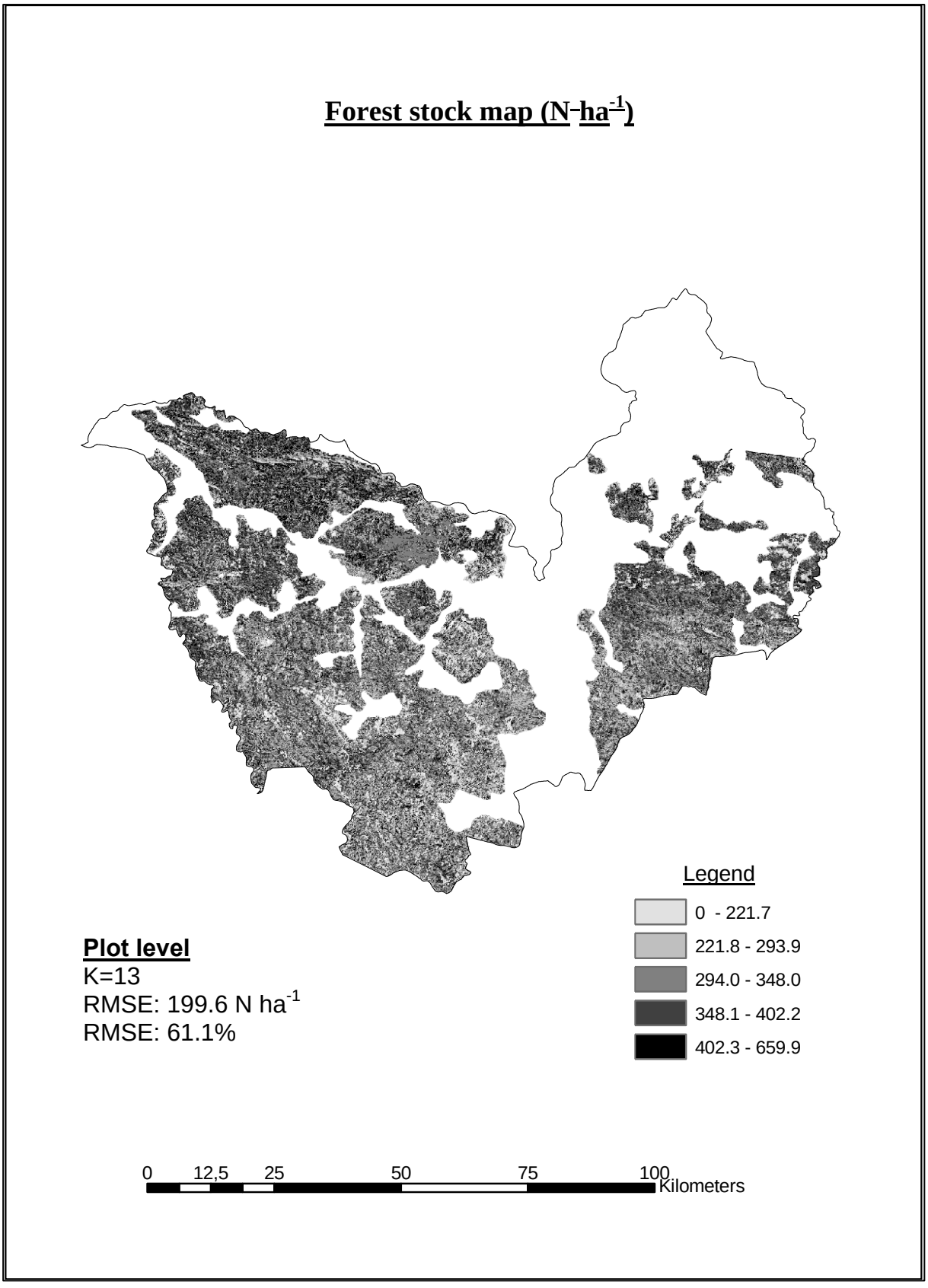
ANNEX 3 – Commercial volume mapping (plot level)



ANNEX 4 – Commercial volume mapping (pixel level)



ANNEX 5 – Forest stock mapping (plot level)



ANNEX 6 – Forest stock mapping (pixel level)