# Methods for estimating volume, biomass and tree species diversity using field inventory and airborne laser scanning in the tropical forests of Tanzania

Metoder for estimering av volum, biomasse og treslagsdiversitet ved hjelp av feltinventering og flybåren laserskanning for tropiske skoger i Tanzania

Philosophiae Doctor (PhD) Thesis

Ernest William Mauya

Department of Ecology and Natural Resource Management
Faculty of Environmental Science and Technology
Norwegian University of Life Sciences

Ås 2015



Thesis number 2015:82
ISSN 1894-6402
ISBN 978-82-575-1317-7

**PhD Supervisors:**

Professor Tron Eid
Department of Ecology and Natural Resources Management
Norwegian University of Life Sciences
P.O. Box 5003, NO-1432 Ås,
Norway

Professor Erik Næsset
Department of Ecology and Natural Resources Management
Norwegian University of Life Sciences
P.O. Box 5003, NO-1432 Ås,
Norway

Professor Terje Gobakken
Department of Ecology and Natural Resources Management
Norwegian University of Life Sciences
P.O. Box 5003, NO-1432 Ås,
Norway

Dr. Ole Martin Bollandsås
Department of Ecology and Natural Resources Management
Norwegian University of Life Sciences
P.O. Box 5003, NO-1432 Ås,
Norway

Dr. Eliakimu Zahabu
Department of Forest Mensuration and Management
Sokoine University of Agriculture
P.O. Box 3013 Morogoro,
Tanzania

**PhD Evaluation Committee**

Dr. Ross Nelson
Hydrospheric and Biospheric Sciences Lab. NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771. (105 Annapolis View Road, Stevensville, MD 21666)
USA

Associate Professor Henrik Meilby
Section for Environment and Natural Resources
Department of Food and Natural Resource Management
Copenhagen University
Rolighedsvej 23, 1958 Fredriksberg C.
Denmark

Professor Ole Hofstad
Department of Ecology and Natural Resources Management
Norwegian University of Life Sciences
P.O. Box 5003, NO-1432 Ås.
Norway

## Acknowledgements

Firstly, I would like to express my special appreciation to my main supervisor Prof. Tron Eid for his guidance, encouragement, and advises throughout the study period. In a more special way, I would like also to extend my heartfelt thanks to my co-supervisors Prof. Erik Næsset, Prof. Terje Gobakken, and Dr. Ole Martin Bollandsås, for introducing me to the field of airborne laser scanning and for their valuable contributions in the process of manuscripts development and write up of the thesis. My sincere appreciation goes also to my co-supervisor from Sokoine University of Agriculture (SUA), Dr. Eliakimu Zahabu for his guidance during the fieldwork in Tanzania and support in the whole period of the study.

Beside my supervisors, I would like to thank Dr. Liviu Theodor Ene for his daily encouragement in a more brotherhood relationship together with insightful comments on statistics, but also for the hard questions which helped me to widen my research skills in various perspectives. I am also indebted to Prof. Rogers Malimbwi from SUA, for his encouragement and moral support, but more important for bringing me to the realm of forest inventory by hiring me in the first place as a research assistant back in 2005.

The financial support of this PhD work has been granted by the project entitled "Climate Change Impacts, Adaptation, and Mitigation in Tanzania" supported by the government of Norway. I am thankful to the project coordinators, in particular Prof. Lars Olav Eik and Prof. Salim Maliondo, for their supports throughout the study period. Much appreciations also goes to my employer SUA for granting me a four years study leave and for providing necessary infrastructure during my field work and holiday when I was in Tanzania. In particular, I am grateful to my immediate boss Dr. Dos Santos Silayo for ensuring that I received all the necessary supports.

I would like also to thank my current and former fellow PhD-students and administrative staffs at INA for creating an open and good learning and working environment. In particular, I am grateful to Roar, Deo, Beatrice, Severin, Endre, Johannes, Daud, Victor, Greyson, Meley, Samora, Asmelash, Stig, Espen, Grethe, Mette, Ole Wiggo, and Ågot. I have also enjoyed being attached to the SkogROVER research group, which has been socially and academically rewarding.

Special thank goes also to my brothers and sisters of the William Hamis Mauya family for all of the sacrifices that they made on my behalf. You were always nearby whenever I needed your assistance, despite the geographical distance. I cannot also express my gratitude to my mother

Chiristina in words. her unconditional love to me has been my greatest strength. She had always asked 'Baba utakuja lini? meaning that. 'My son. when are you coming back?' now is over. I will be coming home soon. To mother and all my brothers and sisters. I say '*Hongeai sana*' (Sambaa). '*Ahsanteni sana*' (Swahili). meaning thank you very much.

My sincere appreciation also goes to my wife Hanney and daughter Erika for being so patient and understanding. They gave me freedom to do my work. though it was hard for them to stay away from me. Finally yet importantly. I thank God. the almighty father for granting me the capability to reach this goal.

This PhD thesis is dedicated to my late farther William Hamis Mauya. who devoted his resources to lay down the foundation of my education.

Ås. 28-08-2015.

Ernest William Mauya

## Contents

**Abstract**

Deforestation and forest degradation in the tropical countries have reduced the extent of forest and woodlands, which conserve biodiversity, provide essential resources to people and help in mitigating climate change through carbon sequestration. Forest conservation projects need methods for estimating tree species diversity to effectively generate information necessary for implementing biodiversity management plans, while greenhouse gas reduction programmes such REDD+ (Reducing Emissions from Deforestation and Forest Degradation) require robust methods to estimate volume and aboveground biomass (AGB). Such methods are also needed in the context of general forest management planning. The four papers included in this thesis are aimed to test and evaluate methods for estimating volume, AGB, and tree species diversity using field and remotely sensed data in the tropical forests and woodlands of Tanzania. In paper I, tree models for estimating total, merchantable stem, and branch volume applicable for the entire miombo woodlands of Tanzania were developed. In Paper II, III, and IV the potential of airborne laser scanning (ALS) data for predicting AGB and measures of tree species diversity was tested and evaluated. The results have shown that ALS data can be used for predicting AGB with reasonable accuracy by using both parametric and non-parametric approaches. Effects of plot size on the AGB estimates were investigated and the results indicated that the prediction accuracy of AGB in ALS-assisted inventories improved as the plot size increased. Finally, the results showed that measures of tree species diversity and particularly tree species richness and Shannon diversity index, can potentially be predicted by using ALS data.

## List of papers

This PhD thesis is based on the following papers which are referred to by their roman numerals (I-IV).

**Paper I**

Mauya, E.W., Mugasha, W.A., Zahabu, E., Bollandsås, O.M., Eid, T., 2014. Models for estimation of tree volume in the miombo woodlands of Tanzania. Southern Forests: a Journal of Forest Science 76, 209-219.

**Paper II**

Mauya, E.W., Ene, L., Bollandsås, O.M., Gobakken, T., Nasset, E., Malimbwi, R., Zahabu, E., 2015. Modelling aboveground forest biomass using airborne laser scanner data in miombo woodlands of Tanzania. Carbon Balance and Management (under review).

**Paper III**

Mauya, E.W., Hansen, E., Gobakken, T., Bollandsås, O.M., Malimbwi, R., Næsset, E., 2015. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. Carbon Balance and Management 10, 1-14.

**Paper IV**

Mauya, E.W., Bollandsås, O.M., Eid, T., Gobakken, T., Næsset, E., 2015. Modelling and predicting measures of tree species diversity using airborne laser scanning data in miombo woodlands of Tanzania (manuscript).

## 1.0. Introduction

### 1.1. Tropical forests and REDD+

Tropical forests account for about 44% of the global forest area (McMahon, 2014). They encompass various forests types including rainforests, mangroves, montane forests, dry forests, and wooded savanna systems (woodlands) (Brandon, 2014; FAO, 2001; Lewis, 2006). Tropical forests are critical to local, regional, and global climate cycles principally through the moisture and carbon that they store, where approximately 271 ± 16 Pg C (Pg, 1Pg = 1.0$^9$ tons) (Grace et al., 2014) is stored in the tropical forests. Beyond carbon storage, tropical forests also contain higher levels of biodiversity than any other type of forests on the planet, holding 2/3 of land-based species (Brandon, 2014). Additionally, these forests provide a wide range of ecosystem services including timber, fuelwood, water purifications, and they have cultural and religious values as well. These benefits are crucial to the more than 50 million people who live in tropical forests and many millions of others who are indirectly dependent on the services from these forests (Hewson et al., 2014). Despite their potential, tropical forests are threatened by deforestation and forest degradation, mainly caused by human induced activities such as timber and fuelwood extraction, conversion of the forest to land uses for agriculture farmland, oil and gas production, mining, and infrastructure development (Lanly, 2003; Venter and Koh, 2012). Loss of biodiversity and increase in global carbon emissions are the major consequences of deforestation and forest degradation in the tropics, which in turn possess a threat to future global climate stabilization. It is estimated that between the years 1990 and 2010 land uses in the tropics including both deforestation and forest degradation, have emitted about 1.4 Pg C per year which is equivalent to 15% of the global human induced carbon emissions (Houghton, 2013).

To address the concerns over the conservation of tropical forest and to mitigate adverse effects of carbon emissions on global climate change, a compensation-based policy mechanism to reduce emission from deforestation (RED), was firstly introduced to United Nations Framework Convention on Climate Change (UNFCCC) in 2005, at the 11th Conference of Parties (COP) in Montreal (Wertz-Kanounnikoff and Kongphan-apirak, 2009). Later in 2007, at the 13th COP held in Bali, RED was expanded to include emissions from forest degradation and hence became REDD. In 2008, at the 14th COP in Poznan, REDD was finally expanded to REDD+ by including aspects related to forest conservation, sustainable management, and enhancement of forest carbon stock, thereby marking the addition of the "plus" to REDD, i.e. REDD+ (Birdsey et al., 2013).

Under the REDD+ mechanism, participating developing countries receive financial incentives (i.e. payments) for their verified success (i.e. performance) in reducing carbon emissions from forest-related activities as well as enhancing the removal of the carbon from the atmosphere (i.e. enhancing the carbon stock) (Goetz et al., 2015). This mechanism has been accepted as a low-cost and promising approach for mitigating climate change (Angelsen and Brockhaus, 2009) that also will secure many ecological functions of forests, including biodiversity conservation and provision of a number of ecosystem services. The interest among developing countries to prepare for hosting REDD+ projects, and in testing the potential mechanisms, has increased significantly since the initial discussions under UNFCCC in 2005.

In spite of the increasing attention drawn to the REDD+ policy, and the large number of pilot projects currently being implemented in the countries, several key aspects of the REDD+ policy have not yet been fully resolved in many tropical countries. They include, among others, development of baseline scenarios (i.e. reference levels against which the enhanced storage of carbon can be measured), and designing credible and efficient forest carbon Measuring, Reporting, and Verification (MRV) systems (Mattsson et al., 2012). Such aspects are fundamentally important because payments for carbon offsets (i.e. financial benefits) under the REDD+ mechanism rely heavily on the reliable estimates of forest carbon stocks and changes over time (Goetz et al., 2015).

Development of MRV systems for REDD+ implementation requires accurate methods for estimation of above ground biomass (AGB) as the primary variable for computing carbon stored in forests. According to Lu (2006), estimation of AGB can be done by using methods based on (1) field measurements, (2) geographical information systems (GIS), and (3) remote sensing. Field based approaches are the most accurate methods for estimation of AGB. However these methods are often time consuming, labor intensive, difficult to implement in the remote areas and they are impossible to apply for large geographical areas with reasonable cost and precisions (Lu et al., 2014; Saarela et al., 2015). GIS-based methods are also difficult because of problem in obtaining good quality ancillary data (Lu, 2006; Lu et al., 2014). On the other hand, remote sensing-assisted methods have gained a wide acceptance for AGB estimations, given their ability to account for limitations related to sample size, time lines, expenses, and accessibility. Moreover, remotely sensed data can provide a synoptic view over large areas and greatly increase the efficiency and usefulness of limited conventional field-based methods (Patenaude et al., 2005; Sinha et al., 2015). Remotely sensed data are therefore considered to play a fundamental role in the development of cost-efficient methods for AGB estimations needed in REDD+ MRV systems (Asner et al., 2012a;

GOFC-GOLD, 2011). Thus it is important to understand and quantify the contribution of different remotely sensed data on improving the estimates of AGB under different forest conditions. This may guide investment decisions in development of cost-efficient MRV systems in the tropical countries. Remotely sensed data have also been reported to be useful for monitoring and assessing different aspects of forests biodiversity, including tree species diversity (e.g. Leutner et al., 2012; Oldeland et al., 2010). In light of the current need for assessment of the impact of REDD+ on biodiversity conservation (Ehara et al., 2014; Vanderhaegen et al., 2015), remotely sensed data may provide valuable information for biodiversity assessment at spatial scales that hardly can be provided by conventional field-based methods at reasonable costs.

Despite the potential of the remote sensing-based methods for REDD+ related issues including sustainable forest management and biodiversity conservation, there has been scarcity of relevant studies in relation to tropical forests, particular in East Africa. Because of that, the Tier-3 approach to carbon inventory that was proposed by Intergovernmental Panel on Climate Change (IPCC) (Eggleston et al., 2006) and further elaborated by Gibbs et al. (2007) can be challenging to implement in many tropical contries in Africa. However, data from airborne laser scanning (ALS) - one among the most promising remote sensing techniques for precise estimation of AGB and other forest properties, have potentials for successful application to tropical forests.

### 1.2. Status of forest resources and REDD+ MRV system in Tanzania

Tanzania main land has a total land area of about 88.3 million ha (MNRT, 2015). Forests and woodlands occupy about 48.1 million ha, equivalent to 55% of the total land area. About 44.7 million ha, or 92% of the total forest area, is classified as woodlands, out of this miombo woodlands make up by far the largest part. The remaining forest areas are mainly occupied by tropical high (mountain) forests, lowland forests, mangrove, and plantation forests . Like in other tropical countries, deforestation and forest degradation are the major challenges for management of forest and woodland resources in Tanzania (Mbwambo et al., 2012). It is estimated that Tanzania lost an average of 403,350 ha or 0.97% of its forest cover per year from 1990 to 2010 (FAO, 2010a). This was mainly a result of heavy pressure from agricultural expansion, livestock grazing, wild fires, and general over-exploitation of wood resources due to different human activities (Blomley and Iddi, 2009; FAO, 2010b). The Tanzanian government has implemented several initiatives to address the challenges of deforestation and forest degradation through adoption of legal frameworks and implementation of participatory forest management regimes (Mbwambo et al., 2012; Treue et al., 2014). The main sources of finance for these initiatives have been obtained from charges levied on the major forest products and services, state budget allocation to the local forestry administrations, and grants obtained from development partners. However, in the recent decade limited financial resources are compelling the country to identify innovative financing mechanisms to support forest management activities outside these traditional channels (URT, 2012).

The emergence of REDD+ under UNFCCC has therefore been considered as an exceptional opportunity for the Tanzanian government to obtain financial resources that will help to improve the management of the forest resources by reducing deforestation, forest degradation, and loss of biodiversity (Araya and Hofstad, 2014; Zahabu et al., 2007). To date Tanzania has already participated in the REDD+ readiness mechanism of the United Nations Collaborative Program on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN REDD), which aimed to enhance government capacity to coordinate, implement, and monitor the REDD+ process (Chiesa et al., 2009). The Tanzanian government also in 2008 signed a letter of intent with the Norwegian government to establish a partnership to address climate change. This partnership intends to build capacity and create carbon accounting methodologies, as well as to support projects that directly seek to reduce deforestation and forest degradation. As a result, Tanzania has become a pioneer among the REDD+ countries with a larger number of sub-national REDD+ projects than any other African country (Lin et al., 2014).

Additionally, Tanzania, with financial support of the Government of Finland and the FAO-Finland Forestry Programme, between 2009 and 2014 established and conducted a national forest inventory, namely the National Forestry Resources Monitoring and Assessment (NAFORMA), with permanent sample plots distributed all over the country (MNRT, 2015). NAFORMA sample plots forms the backbone of a system for estimation of the current state of the forest resources needed for national forest policies and general forest management as well as inputs for monitoring changes in carbon stock needed for the development of national carbon MRV systems under REDD+ implementation. NAFORMA data may also be used for biodiversity monitoring since much information related to biodiversity have been collected as part of the field campaign. This may generate an important platform for assessment of the impact of REDD+ on enhancing biodiversity conservation and other related policies that are aimed towards conservation planning.

### 1.3. Rationale for the study

Irrespective of the well updated national forest inventory data (i.e. NAFORMA), reliable information on forest resources in Tanzania will in the future heavily depend on the availability of accurate methods for quantification of different forest attributes including volume, AGB, carbon stock, and tree species diversity. For volume, AGB, and carbon stock, relevant tree models covering different vegetation types of Tanzania are needed for computation of plot-based estimates, which is a fundamental part of any forest inventory. However, such models are often limited in Tanzania. Besides that, for development of cost-efficient forest monitoring system needed for REDD+MRV, conservation planning, and general forest management decision-making, methods based on remote sensing techniques are needed for enhancing the precisions of NAFORMA plot-based estimates (MNRT, 2015; Tomppo et al., 2014). Unfortunately, to date empirical evidence on the accuracy and performance of remote sensing-based methods for estimation of different forest attributes in Tanzania are limited. Most of the reported remote sensing-based studies have focused on mapping and quantification of forest cover changes using satellite-based remote sensing (Lupala et al., 2015; Swetnam et al., 2011; Tabor et al., 2010), with little attention on estimation of forest attributes, like those mentioned above. Furthermore, more recently developed techniques such as ALS, which has gained widespread application in forests in other parts of the world, has to date not been validated for forest attribute estimation purposes in Tanzania. The current thesis aims at filling some of this gap.

## 2.0. Research objectives

The main objective of this study was to test and evaluate methods for estimating forest resources as a basis for forest management decision-making and development of a REDD+ MRV system in Tanzania. The thesis is divided into the following sub-objectives addressed in four different papers:

1. To develop allometric models for estimating tree volume in miombo woodlands of Tanzania (Paper I).
2. To model the relationship between AGB and ALS data in miombo woodlands of Tanzania and assess the model performance (Paper II).
3. To assess effects of field plot size on prediction accuracy of AGB in ALS-assisted inventories in tropical rainforests of Tanzania (Paper III).
4. To model and predict measures of tree species diversity using ALS data in miombo woodlands of Tanzania (Paper IV).

## 3.0. Background

### 3.1. Analytical conceptual framework of the thesis

Accurate information on forest resources, needed for making informed decisions on forest management as well as for development of REDD+ MRV systems, requires application of reliable methods for estimation of different forest attributes. Field data and remotely sensed data are the primary data sources for development of such methods. In the conceptual framework presented in Figure 1, data on stem and branch volume of trees derived from destructive sampling are obtained from the field data. Such data are used as an input to develop allometric tree volume models using ordinary least square (OLS) and nonlinear least square (NLS) regression techniques. Field data also provided plot-based information on AGB in Mg per ha, AGB in Mg per ha on different field plot sizes, and measures of tree species diversity. The information is combined with ALS metrics (derived from ALS data) using different statistical methods including: OLS, linear mixed effects models (LMMs), and k-nearest neighbors (*k*-NN). The outputs are methods in terms of prediction models based on the ALS data for estimating AGB and measures of tree species diversity. These models in combination with allometric tree volume models form methods that can be used for estimation of forest resources needed for making informed decisions in forest management and for development of a REDD+ MRV system.

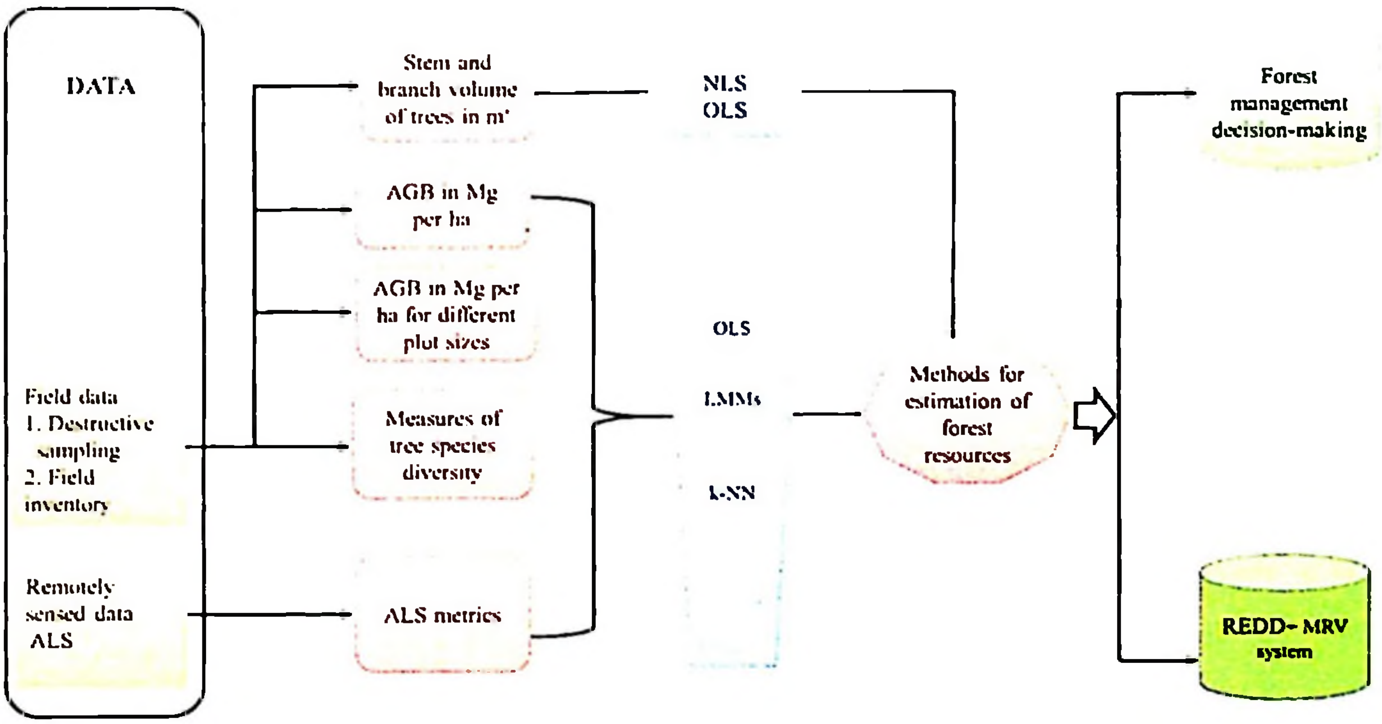



Figure 1. Analytical conceptual framework for the thesis.

### 3.2. Tree AGB and volume models

Allometric tree AGB and volume models (models for individual trees) are crucial for quantifying many forest products and services such as commercial timber volume, bioenergy volume, and carbon stock. Development of tree AGB models usually involves harvesting and weighing a sample of trees across a range of diameter and/or height classes, and then estimating the parameters of a model relating AGB to diameter at breast height (*dbh*) or a combination of *dbh* and total tree height (*ht*) using regression techniques. Such models can then be used to estimate AGB using *dbh* or combinations of *dbh* and *ht* measured in forest inventories. Another approach is to estimate AGB based on volume and subsequently use an expansion factor (Birdsey et al., 2013). The latter approach of using tree volume models and expansion factors has been used by most of the countries for reporting national forest biomass (FAO, 2010a), as compared to the direct use of tree AGB models.

The process of developing tree AGB and volume models is time consuming and expensive as it relies on destructive sampling as previously mentioned. In tropical forests, where there is a wide range of different tree species, development of allometric tree AGB and volume models valid for large geographical areas has always been a challenge. For dry tropical forests and woodlands, relatively few models for estimating tree AGB and volume have been developed as compared to moist and wet tropical forests (Henry et al., 2011). For the miombo woodlands of Tanzania, Malimbwi et al. (1994), Chamshama et al. (2004), and Mwakalukwa et al. (2014) have developed volume and AGB models. The applicability of these models is, however, limited by the sample size and tree size ranges of the data used for modelling. Furthermore, the models cover only limited geographical ranges of miombo woodlands and thus none of them can be regarded as general models covering the whole country. This is obviously a challenge for the development of a sustainable forest management and an effective forest monitoring system needed for implementation of a REDD+ policy, irrespective of the availability of more recently updated forest inventory data in the country. Thus, robust tree models in terms of geographical and tree size coverages are needed both for REDD+ reporting and for supporting the development of other decision-support tools for forest management (Mugasha, 2014). General tree AGB models for miombo woodlands covering the entire miombo woodlands of Tanzania have recently been developed by Mugasha et al. (2013), but no general tree volume models exist. The main objective of Paper I was therefore to develop tree volume models for miombo woodlands of Tanzania. The specific objectives were to (1) develop both general and site specific models, (2) develop models for total volume, merchantable stem volume, and branch volume, and (3) compare the performance of the general models with models developed previously by other researchers.

### 3.3. Remote sensing support for forest inventory

As part of the requirements for development of a cost effective REDD+ MRV system, the application of remote sensing techniques has been considered as a relevant option for improving the statistical precision of AGB estimates (GOFC-GOLD, 2011). Likewise, for deriving information needed for forest management at different spatial scales, ranging from local to national levels. Types of remote sensing data that have been used for estimation of AGB include optical, synthetic aperture radar (SAR), and ALS (De Sy et al., 2012). However, the performance of optical remote sensing, and to a lesser extent SAR, in tropical forests are affected by forest structure complexity, canopy density, and cloud coverage (Gibbs et al., 2007). As a result of this, both of the two remote sensing techniques saturate at lower AGB levels, say, around 20-250 Mg $ha^{-1}$ of AGB (Kaasalainen et al., 2015; Patenaude et al., 2005), although SAR data have shown to saturate at higher values as compared to optical remote sensing (Lucas et al., 2007; Sinha et al., 2015).

ALS has proved to overcome some of the limitations of other remote sensing techniques, because of its ability to retrieve three-dimensional (3D) forest structures at high spatial resolution. Such information is more useful for forest inventories as well as for ecological applications than the information from any of the other remote sensing techniques (Coops et al., 2004; Lefsky et al., 2002; Vauhkonen et al., 2014). The strengths of ALS in estimating forest attributes have been investigated from both research perspective and operational perspective, and ALS is currently used in operational management inventories in the Nordic countries (Vauhkonen et al., 2014). Promising results from the tropical forests of Asia and south America have also been reported (e.g. Asner et al., 2012b; Hou et al., 2011; Mascaro et al., 2012). In the tropical forests of Africa, the application of ALS is at very early stages despite its wide acceptance as a potential tool for AGB estimations necessary for the development of a MRV system under REDD+ initiatives. In Tanzania, this thesis is one among the earlier works to test the potential of ALS for estimating AGB.

#### 3.3.1. Modelling and predicting AGB using ALS data

Application of ALS for estimation of forest attributes including AGB has commonly been done by using two methods, namely; the individual tree-based approach and the area-based (ABA) approach (Hyyppä et al., 2008; Vauhkonen et al., 2014). When using the individual tree-based approach, individual trees are detected and tree-level attributes, such as AGB, height, and volume, are measured or predicted from the ALS data, i.e. the basic unit is an individual tree. The potential of the individual tree-based approach has been demonstrated in different studies (e.g. Hauglin et al., 2012; Maltamo et al., 2004; Vauhkonen et al., 2010) in which forest attributes such as tree position, AGB, tree height, and volume have been accurately predicted. However, the individual tree-based approach requires denser ALS data, which then increases the ALS acquisition cost (Hyyppä et al., 2008). Furthermore, with the individual tree-based approach, linking a tree crown delineated from ALS data to a field-measured tree, requires positions of the tree measured in field. This might be a limitation in a complex ecosystem such as tropical forests, where the number of trees per unit area usually is larger than in temperate forests (Lu et al., 2014). In such a situation, the ABA is the most favored alternative. With the ABA, various metrics are extracted from the ALS data recorded on the field plot, and statistical models relating the ALS metrics and plot level forest variables are constructed. The models are then used to predict for example AGB for all areas covered by ALS data (Næsset, 2014).

A "wall-to-wall" ABA is commonly used in operational forest inventories in the Nordic countries (Næsset, 2014), mainly because of lower cost and maturity of the approach as compared to the individual tree-based approach. However, for larger territories such as entire regions (districts) or even nations, it is currently not economically feasible, and most likely not required from an accuracy perspective, to provide wall-to-wall data. It has therefore been proposed to use just a sample of ALS data collected along some selected flight lines over the territory of interest (McRoberts et al., 2014). These flight lines should primarily be distributed over existing ground plots according to sound statistical principles, allowing the development of models that tie the AGB on the ground to the ALS data. These models are then used to predict AGB over the entire area covered by ALS strips, and subsequently these predictions are used for final estimation of AGB for the area of interest using either design-based model-assisted or model-dependent inferential frameworks (e.g. Gobakken et al., 2012; Gregoire et al., 2010). Thus, the quality of the AGB estimates produced by ALS-based inventories relies heavily on the development and application of appropriate predictive AGB models.

Many statistical techniques have been tested for predicting AGB when using ALS data and they can be grouped into two broad categories: parametric and non-parametric. Each method has its own strengths and requirements for data inputs that have created challenges in identifying the optimal method for improving prediction accuracy of AGB when using ALS data (Lu et al., 2014). Because of that, comparative analysis of different methods such as OLS, random forests, and nearest neighbor techniques have been conducted in temperate and boreal forests (e.g. Bollandsås et al., 2013; Gagliasso et al., 2014; Gleason and Im, 2012). All these efforts were aimed at improving performance of ALS-aided inventories in terms of prediction accuracy of the models. Unfortunately, the majority of these studies have mainly focused on temperate and boreal forests as indicated above – forests that may differ quite much in structure, composition, and internal distribution of AGB between different tree and canopy elements compared to tropical forests. In the tropical forests, particular those of Africa, the reported studies have so far focused on the use of parametric methods (e.g. Asner et al., 2012b; Hansen et al., 2015), and neither of them has been conducted in the miombo woodlands of Tanzania. The main objective of Paper II was therefore to address this knowledge gap by assessing the performance of parametric and non-parametric methods for modeling and predicting AGB using ALS data in the miombo woodlands of Tanzania. The specific objectives were to (1) compare the performance of parametric and non-parametric methods in modelling and predicting AGB using ALS data, and (2) assess the effects of post-stratification on the prediction accuracy of the parametric models.

### 3.3.2. Effects of field plot size on prediction accuracy of AGB in ALS-assisted inventories

Previous studies in temperate, boreal, and tropical forests have indicated that field plot size is an important sampling parameter that should be taken into account when designing ALS-assisted inventories (e.g. Frazer et al., 2011; Mascaro et al., 2012; Næsset et al., 2015). Model prediction errors have been reported to decline with increase in plot size, although the ranges of the plot sizes were varying from one forest type to another. For example, plots sizes ranging from 0.02 ha (Næsset et al., 2011) to 0.3 ha (Yao et al., 2014) have been reported from temperate and boreal forests, while for tropical forests plot sizes up to 1 ha have often been recommended (e.g. Asner and Mascaro, 2014; Mascaro et al., 2011; Zolkos et al., 2013). In all cases the studies indicated that an increase in plot size resulted in better prediction accuracy because of reduction of the errors caused by the discrepancies between ground and ALS-based measurements (Zolkos et al., 2013).

Given the potential of ALS for development of REDD+ MRV systems, it is important to determine the plot size ranges that should be considered for designing future ALS-assisted inventories and to address MRV standards in different forest types in tropical forests. A field plot size equivalent to a radius of 15 m has commonly been used in field-based forest inventories in Tanzania (Tomppo et al., 2014). For ALS-assisted inventories, however, such a plot size may not result in sufficiently good prediction accuracy due to the challenges discussed above, and especially so in high forests where the potential for discrepancies between large trees measured on the ground (tree stems inside the plot) and ALS measurements of tree crowns (crowns rather than stems inside the vertical extension of a plot) are substantial. Therefore the main objective of Paper III was to assess the effects of field plot size on prediction accuracy of AGB in ALS-assisted inventories in tropical rainforests of Tanzania. The specific objectives were to (1) examine the effects of field plot size on AGB regression model quality, (2) assess plot boundary effect and its impact on model quality based on the field data, and (3) quantify the precision of ALS-assisted estimates of AGB relative to field-based estimates of AGB assuming the same design for different plot sizes.

### 3.3.3. Modelling and predicting measures of tree species diversity using ALS data

ALS has gained wide acceptance also in ecologically-based studies in the recent decades due to its ability to quantify the3D structure of forests, which is of particular interest in characterizing species diversity of different taxonomic groups in the forest (Müller and Vierling, 2014). Several studies have explored the potential of ALS to model the assemblage, compositions and diversity of insects, spiders, and birds (e.g. Muller et al., 2009; Vierling et al., 2011; Vogeler et al., 2014). Measures of tree species diversity, including Shannon diversity index and tree species richness have also been reported to be successfully predicted and classified using ALS data (Leutner et al., 2012). Despite the encouraging results from this body of work, few studies have estimated or analyzed the relationship between measures of tree species diversity and ALS data in tropical forests. Lack of ALS data and complexity of the structure, due to the high diversity of tree species in tropical forests, are among the possible reasons for the low number of such studies in tropical forests. In Tanzania, no previous studies have attempted to use ALS data for modelling and predicting measures of tree species diversity. As stated above, remotely sensed data may provide valuable information for biodiversity assessment at spatial scales that hardly can be provided by conventional field-based methods at reasonable costs. The main objective of Paper IV was therefore to assess if ALS data can be used to predict measures of tree species diversity in the miombo woodlands of Tanzania. The specific objectives were to (1) examine the performance of parametric and non-parametric methods for predicting measures of tree species diversity using ALS data, and (2) assess the prediction accuracy of measures of tree species diversity across different vegetation types.

## 4.0. Materials and methods

### 4.1. Study sites

The study sites were located in five different administrative regions in Tanzania (Figure 2) in order to cover the different objectives of the thesis. Out of the five sites, four were mainly dominated by miombo woodlands (i.e. Manyara, Tabora, Katavi, and Lindi) and one (i.e. Tanga) was mainly dominated by tropical rainforest.

*Miombo woodlands*

Miombo woodlands make up a significant proportion of the forested land in Tanzania. The largest concentration of the miombo woodlands in Tanzania is found in the western and southern part of the country (Abdallah and Monela, 2007). Miombo woodlands also extend to other African countries including Angola, Democratic Republic of the Congo, Mozambique, Malawi, Zambia, and Zimbabwe (Dewees et al., 2010). According to White (1983), miombo woodlands may be divided into two major distinct groups: dry and wet miombo woodlands. Dry miombo woodlands occur in areas receiving less than 1000 mm rainfall annually while wet miombo woodlands occur in areas receiving more than 1000 mm. The vegetation of the miombo woodlands is floristically rich characterized by the overwhelming dominance of *Brachystegia*, *Julbernardia*, and *Isoberlinia* tree species belonging to the Fabaceae (legume) family (Backéus et al., 2006). In mature miombo, these species comprise an upper canopy layer made of 10-20 m high trees and a scattered layer of sub-canopy trees. The understory layer is discontinuous and is composed of broadleaved shrubs and suppressed saplings of canopy trees. A sparse, but continuous herbaceous layer of grasses, forbs, and sedges dominate the ground-layer (Campbell et al., 2007). Miombo woodlands soils are typically acidic, highly leached and low in organic matter (Frost, 1996).

*Tropical rainforest*

Most of the Tanzanian tropical rainforests occurs primarily on mountain slopes and are confined to the Eastern Arc Mountains (EAM) system, which is a chain of mountains that stretches from Makambako, southwest of the Udzungwa Mountains in southern Tanzania, to the Taita Hills in south-coastal Kenya. The most important parts of EAM with rainforests are Southern Pare, West Usambara, East Usambara, Nguu, Nguru, Ukaguru, Uluguru, Rubeho, Malundwe, Uduzungwa, Mahenge, and Matengo (Bjorndalen, 1992). Since these forests are spatially isolated from each other, they have more diverse flora and fauna than many other ecosystems. The most important functions of this type of forests are to serve as watershed catchments areas and soil protection areas. Rainfall in most of these rainforests is heavy, with a short dry season. In this thesis, Amani Nature Reserve (ANR), located in East Usambara, was selected as a site representing the typical tropical rainforest existing in Tanzania. ANR is dominated by trees of genera *Afrosersalisia*, *Allanblackia*, *Celtis*, *Drypetes*, *Ficus*, *Isoberlinia*, *Leptonychia*, *Macaranga*, *Myrianthus*, *Newtonia*, *Parinari*, *Sorindeia*, *Strombosia*, *Syzygium*, and *Tabernaemontana* (Hamilton and Bensted-Smith, 1989). Generally, the forest is denser and is often regarded as one among the tropical rainforests with highest values of AGB (e.g. Marshall et al., 2012; Munishi and Shear, 2004), ranging from 43.2 to 1147.1 Mg ha$^{-1}$ (Hansen et al., 2015) in ANR in particular. Thus, the selection of this site was aimed to demonstrate the typical challenges that would be expected when using ALS in tropical rainforests with high AGB densities.

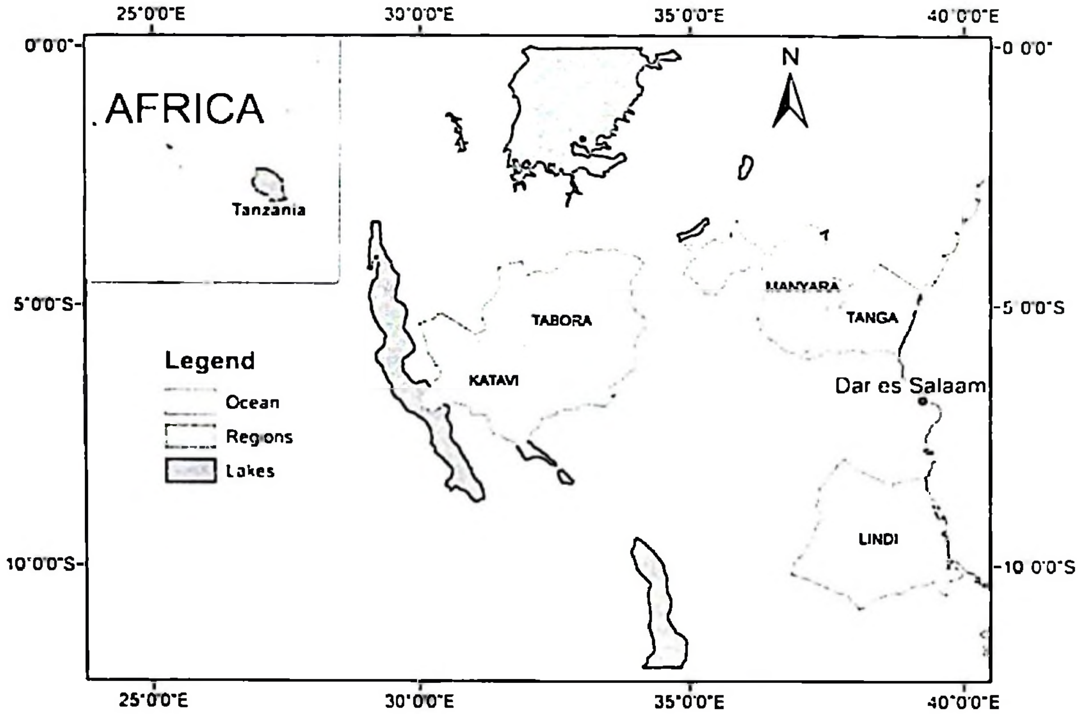



Figure 2. Location of the study sites in five administrative regions of Tanzania.

### 4.2. Data collection

Three types of data, comprising a total of five different datasets, were used in this study, i.e. one dataset based on destructive sampling of trees, two field inventory datasets, and two remotely sensed datasets. Details regarding these datasets are described below.

#### 4.2.1. Destructive sampling data

The dataset based on destructive sampling of trees was used in Paper I. The data were collected from one site in each of the following regions: Manyara, Tabora, Babati, and Lindi (Figure 2) and the selected trees have previously been used to develop general allometric above- and belowground biomass models for miombo woodlands of Tanzania (Mugasha et al., 2013). The sites covered a wide range of growing conditions in the miombo woodlands of Tanzania. To secure an appropriate distribution of sample trees with regard to tree sizes and tree species, information from previous systematic sample plot inventories carried out for each of the four sites was used as guidance when selecting the trees. For each site, sample plots with 15 m radius were established on a systematic grid, and from these plots one or two trees were selected purposely for destructive sampling. Prior to tree felling, all sample trees were identified for species name (local and scientific name), and *dbh* was measured with calipers and diameter tapes while *ht* was measured with Suunto and Vertex hypsometers. A total of 158 trees, including 55 different tree species distributed in the four sites were sampled: Manyara (39), Tabora (40), Lindi (39) and Katavi (40). As part of the destructive sampling procedure, all sample trees were separated into merchantable stem (tree section suitable for timber) and branches including top (to minimum diameter of 2.5 cm).

The volume of each log was calculated by multiplying the cross sectional area at the midpoint of each log with its length. Volume of merchantable stem and branches for a tree was obtained by summing the volumes of the logs of the respective sections for that specific tree. Total tree volume was finally obtained through summation of merchantable stem and branches volumes.

#### 4.2.2. Field inventory data

Two datasets consisting of field inventory data (sample survey plots) were collected. Field inventory dataset 1 was used in Papers II and IV. This dataset was based on the plot locations of the sample plots that were initially established and measured by NAFORMA crews as part of the national forest inventory in Liwale district, Lindi-region, in June 2011. The field work for dataset 1 was conducted eight months later after the completion the initial measurement done by NAFORMA. The aim of the field work was to accurately record the positions of the field plots using high precision positioning equipment and to re-measure the NAFORMA field sample plots in order to have temporal consistence with the time of the ALS data acquisition. Measurements on the plots followed the same protocol as used by NAFORMA in 2011. Field measurements were collected for 65 NAFORMA clusters, each having a total of ten sample plots per cluster (MNRT, 2011). However, in this field work two sample plots for each cluster were omitted because they were outside the corridors (swaths) designated for the ALS data acquisition. Therefore eight plots per cluster were re-measured for the current analysis. On each plot, the (x, y) center coordinates were determined using differential post-processing of dual-frequency Global Positioning System (GPS) and Global Navigation Satellite System (GLONASS) measurements. On each plot, tree species name and *dbh* were recorded following the concentric plot design for the 15 m radius plots described in MNRT (2011). Height measurements were acquired for every fifth tree in the cluster.

AGB for each tree was estimated using allometric models developed by Mugasha et al. (2013). The AGB of the individual trees were then summed to obtain estimates for each plot, which were used in Paper II. The same dataset was used in paper IV to compute measures of tree species diversity, i.e. tree species richness and Shannon diversity index based on the information collected from the tree species names.

Field inventory dataset 2 was used in Paper III. A total of 30 circular field plots were established in ANR and distributed along elevation ranges from 200 to 1000 m above sea level. On each of the 30 plots, all trees with $dbh \geq 5$ cm were callipered and registered for botanic names, local names, and the horizontal distance to the plot center. The horizontal distance was measured from plot center to the front of each tree using a Vertex hypsometer, and half of the tree diameter was added to get the total horizontal distance (i.e. radius). The distance measures were used to generate different plot sizes within the limit of the maximum radius. The maximum radius varied among the 30 plots due to the somewhat unequal performance of the Vertex hypsometer in the different plots (different tree densities etc); 31 m (22 plots), 28 m (2 plots), 26 (1 plot), and 25 m (5 plots). For this study, a minimum radius of 7.98 m and a maximum of radius 30.90 m was used, which is equivalent to the plot sizes 200 $m^2$ and 3000 $m^2$, respectively. Three trees (largest, medium, and smallest) per plot were measured for height using a Vertex hypsometer. Precise field coordinate positions for each plot center were determined by means of the GPS+GLONASS receivers and procedures describe above. AGB was calculated for individual trees within each plot using an allometric model developed from destructively sampled trees in ANR by Masota et al. (2015). The values were finally scaled to per ha basis for each of the plot sizes.

### 4.2.3. Remotely sensed data

Two remotely sensed datasets based on ALS were used in this thesis: Remotely sensed dataset 1 was used in Papers II and IV, while remotely sensed dataset 2 was used in Paper III. Acquisition of the ALS measurements for dataset 1 was done by using a strip sampling approach covering about 26% of the study area in Liwale district. For dataset 2, the ALS measurements were acquired over the entire study area of ANR (i.e. wall to wall). Characteristics of the two remotely sensed datasets are shown in Table 1.

The initial processing of all the ALS data was accomplished by the contractor (TerraTec AS, Norway). Then several ALS metrics consisting of both height and canopy density variables were computed from the ALS data following the procedures described by Næsset (2004) and Gobakken et al. (2012).

Table 1. Characteristic of the remotely sensed datasets.

| Remotely sensed dataset | ALS data characteristics | |
|---|---|---|
| 1 | **Acquisition settings** | |
| | Acquisition date | 10 February to 07 March 2012. |
| | Platform | Cessna 404 |
| | Canopy conditions | Leaf-on conditions |
| | Flying altitude | 1200 m |
| | Flying speed | 77.2 $ms^{-1}$ |
| | **Sensor settings** | |
| | Sensor | Leica ALS-70 |
| | Pulse repetition frequency | 193 kHz |
| | Scan frequency | 36.5 Hz |
| | Pulse density | 1.8 points $m^{-2}$ |
| 2 | **Acquisition settings** | |
| | Acquisition date | 19 to 25 January and 2-18 February 2012 |
| | Platform | Cessna 404 |
| | Canopy conditions | Leaf-on conditions |
| | Flying altitude | 800 m |
| | Flying speed | 70 $ms^{-1}$ |
| | **Sensor settings** | |
| | Sensor | Leica ALS70 |
| | Pulse repetition frequency | 339 kHz |
| | Scan frequency | 58.6 Hz |
| | Pulse density | 10.6 points $m^{-2}$ |

### 4.3. Data analyses

Different statistical techniques and analyses were employed in order to address the objectives of the individual papers.

*Paper I*

OLS and NLS regression techniques were used for fitting the tree volume models. Four model forms were selected and tested based on previous studies and, two of the model forms included *dbh* only as independent variable while the other two included both *dbh* and *ht*. Leave-one-out cross-validation (LOOCV) was used to evaluate the performance of the models. For each model, relative root mean square error (RMSE) and mean prediction error (MPE) were calculated based on the values from the LOOCV. Additionally, pseudo-$R^2$ was computed as an indicator of model fit. Previously developed tree volume models, which have been applied in Tanzania and in the neighboring countries, were also evaluated on our dataset and compared with the developed general model for total tree volume. Site-specific and tree sectional models (i.e. branch and merchantable stem) were also developed and evaluated using the same procedure as for the general model.

*Papers II and IV*

Parametric and non-parametric methods were used to develop statistical models for prediction of AGB and measures of tree diversity using ALS-derived metrics. Specifically LMMs and *k*-NN were used to account for the cluster structure. LOOCV was applied to compare the methods and asses accuracy of the predictions. LOOCV was performed at the cluster level to ensure that the hierarchical data structure was preserved during re-sampling (i.e. leave-one-cluster-out). Effects of post stratification were assessed by fitting separate models for different vegetation and land use types as defined in MNRT (2011).

*Paper III*

Multiple linear regression analysis with OLS was used to develop the statistical models relating the field reference AGB and the predictor variables from the ALS data for each of the plot sizes. In order to assess the performance of the models for each plot size, LOOCV was conducted. RMSE and MPE were used to assess the prediction accuracy across different plot sizes. In addition, the relative efficiency (RE) between ALS-assisted inventory and pure field-based inventory for different plot sizes was calculated as the ratio of the variance estimates of the field-based inventory relative to ALS-assisted inventory. Design-based and model-assisted variance estimators were used. RE values were then used to assess the efficiency gain by using ALS to assist in the estimation for different plot sizes.

## 5.0. Main findings and discussion

### 5.1. Tree volume models for miombo woodlands

In Paper I, tree volume models with *dbh* as independent variable and those that incorporate both *dbh* and *ht* were developed. The pseudo-$R^2$ of the general model with *dbh* only was 0.87 while for the model with both *dbh* and *ht* it was 0.88. Results from the LOOCV indicated that the RMSE value for the model with *dbh* and *ht* was lower as compared to the model with *dbh* only. Based on the results from the LOOCV, the model with both *dbh* and *ht* was considered as the best general tree volume model for miombo woodlands. The RMSE value of this model was 47.6% of the mean value and the MPE value was -0.6%. Both the pseudo-$R^2$ and the RMSE values were in line with most previous studies in the miombo woodlands and related vegetation types in Tanzania and nearby countries (e.g. Abbot et al., 1997; Chamshama et al., 2004; Malimbwi et al., 1994; Mwakalukwa et al., 2014). Site-specific models were also developed with pseudo-$R^2$ ranging from 0.77 to 0.95. The MPE values for the site-specific models were relatively small as compared to the values obtained when the general model was applied for the individual sites. Since the site-specific models performed best for their respective sites, it would be more meaningful to apply the site-specific models for local inventories at the respective sites as recommended by Abbot et al. (1997) and Mugasha et al. (2013). For large-scale inventories, such as national forest inventories (e.g. NAFORMA), the general models should be considered. Assessment of the performance of the previously developed models on the current dataset used for modelling, showed that the model with *dbh* only as the independent variable developed by Malimbwi et al. (1994) significantly under-predicted volume by 8.5%, while the models with *dbh* and *ht* as independent variables developed by Malimbwi et al. (1994) and Chamshama et al. (2004) significantly under-predicted volume by 20.8% and 31.2%, respectively. The model developed by Abbot et al. (1997) significantly over-predicted the volume by 30.6%. The trends of over- and under-prediction are also seen in Figure 3. The observed trends of over- and under-prediction illustrate the importance of being cautious when applying the models beyond geographical and size ranges used to construct the model.

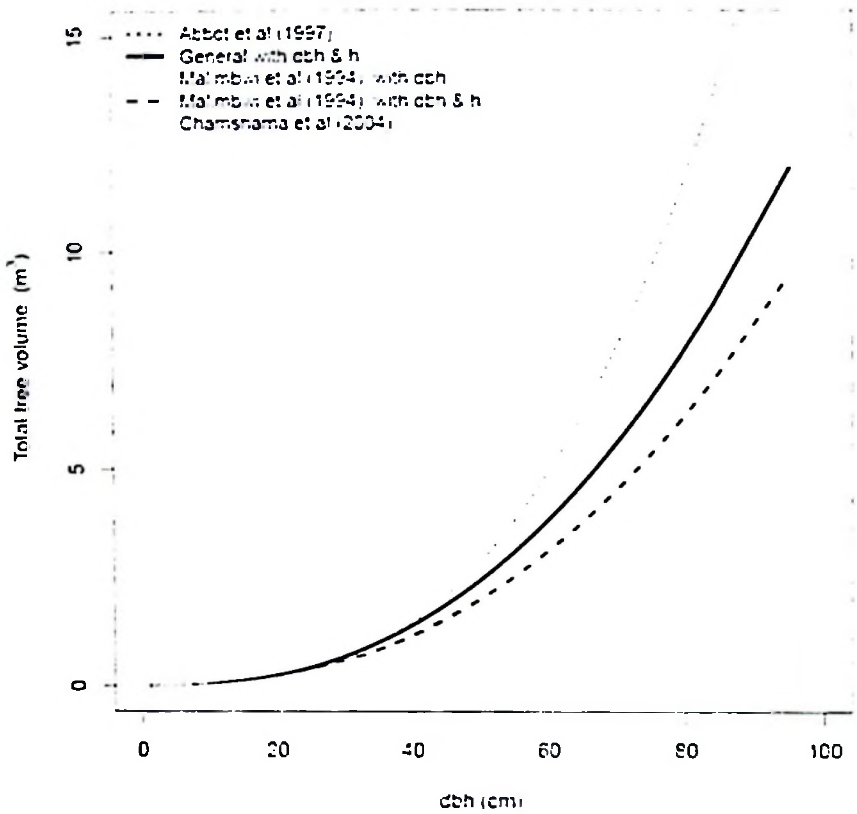



Figure 3. Display of predicted total tree volume over *dbh* based on the general model with *dbh* and *ht* as independent variables developed in this study and based on the models developed by Abbot et al. (1997), Malimbwi et al. (1994), and Chamshama et al. (2004).

### 5.2. Modelling AGB using ALS data

In Paper II, the use of ALS data for modelling and predicting AGB using parametric and non-parametric methods was demonstrated. The results showed that the LMM, with a variance function that accounts for the cluster effects, resulted in the best prediction model ($R^2 = 0.68$, RMSE= 46.8%). The *k*-NN imputation method was also tested with different *k* values. The imputation with a *k value* of 10 turned out to be the best with an RMSE value of 55.9%. Even though the results suggest that there were only small differences in prediction accuracy for the two methods, they can both be considered for AGB prediction and estimation using ALS data in the miombo woodlands. The accuracy of the methods in terms of the reported $R^2$ and RMSE from the LOOCV were in line with most of the previous studies reported from tropical forests (e.g. Asner and Mascaro, 2014; Asner et al., 2012b; Hansen et al., 2015). Post-strata models fitted using LMMs performed well with relatively smaller RMSE values as compared to the RMSE values obtained when the non-post-stratified model was applied to different post strata. However, use of post-strata models for operational prediction purposes would require thematic maps for the land use classes and vegetation types in order to know where to apply different stratum-specific models. Such maps were not available at the time when the analyses were conducted. Thus, since the difference between the non-post-stratified model and the post-strata models were modest, the non-post-stratified model (which disregards the land use and vegetation types) is considered to be more adequate for most applications that will involve large-scale AGB estimation supported by ALS data, at least until high-quality thematic maps are made available. Despite the fact that both of the non -post-stratified and post-stratified models resulted in reasonable prediction accuracies, there were some limitations that may have reduced the prediction accuracy of the models. The plot size used in this study was relatively small as compared to the most widely used field plot sizes in ALS-based studies from the tropical forests (e.g. Asner and Mascaro, 2014; Asner et al., 2012b; Mascaro et al., 2011). Smaller plots are considered as a source of poor prediction accuracy because of the increase in the errors associated with the mismatch between the ground field measurements and ALS measurements (Frazer et al., 2011; Zolkos et al., 2013). The effects of plot size on ALS based inventories are further discussed in Paper III.

**5.3. Effects of field plot size on prediction accuracy of AGB in ALS-assisted inventories**

In Paper III, increasingly stronger relationships were found between field reference AGB and the ALS-derived metrics as the plot sizes increased from 200 $m^2$ to 3000 $m^2$. The adjusted $R^2$ increased from 0.35 to 0.74, while the relative RMSE decreased from 63.6 to 29.2%, indicating that the prediction accuracy improved with increasing plot size (Figure 3). The quality of the model fit as indicated by adjusted $R^2$ and RMSE (Figures 3 and 4) are in line with most of the previous studies reported from the tropical forests (e.g. Asner et al., 2012b; Hansen et al., 2015), although direct comparison should be done with caution due to the differences in factors such sample sizes, forest structure, and geographical range of the data used in the reporting. Results from the variance estimations further indicated that the relative efficiency of the ALS-assisted estimates was improved when the plot sizes increased compared to field based estimates. The relative efficiency values were all larger than one with a maximum value of 7.7 for a plot size of 3000 $m^2$, indicating that ALS-assisted estimates is about seven times more efficient than field-based estimates when using a plot size of 3000 $m^2$ in this forest type. Under simple random sampling a relative efficiency of, say, 7, translates to a need for seven times as many field plots for pure field-based estimation to provide the same precision of the AGB estimates as an ALS-assisted estimation. The gain in relative efficiency was more pronounced as plot size increased, suggesting that from a pure technically perspective larger plots are more favorable when ALS-data are used to assist in the estimation. It should be noted though that differences in costs associated with different plot sizes - and thus cost effectivity - were not considered in this study. Numerous causes for the seemingly higher relative efficiency of ALS-assisted estimates with larger plots can be given: reduced circumference to area ratio, spatial averaging, and smaller effects of positioning errors are among the possible causes (e.g. Frazer et al., 2011; Næsset et al., 2015; Zolkos et al., 2013).

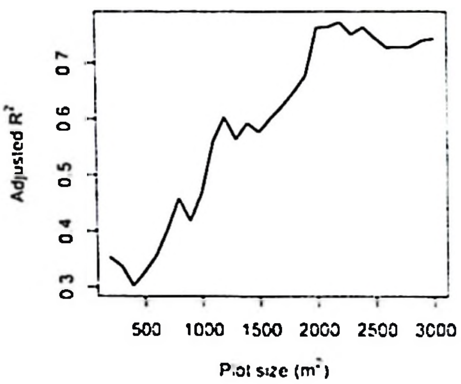



Figure 4. Adjusted $R^2$ versus plot size.

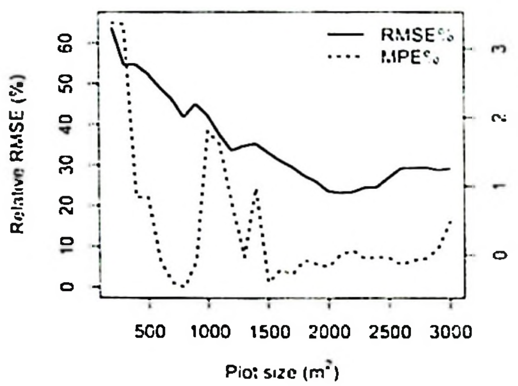



Figure 5. Relative mean prediction error (MPE%) and relative root mean square error (RMSE%) as percentage of field reference AGB versus plot size.

**5.4. Modelling measures of tree species diversity using ALS data**

In Paper IV, tree species richness and Shannon diversity index were modeled using ALS data. The explained variations (i.e. $R^2$) of LMM for tree species richness were relatively high as compared to Shannon diversity index. This indicates that the ALS metrics explain more of the variation in tree species richness as compared to Shannon diversity index. Results from the LOOCV showed that the RMSE values for both tree species richness and Shannon diversity index were slightly higher when using *k*-NN as compared LMMs (Table 2). However, both methods could be considered for prediction and estimation of measures of tree species diversity using ALS data. Prediction accuracy of both methods varied across different vegetation types.

The major findings in terms of model quality criteria, such as $R^2$ and RMSE%, are in accordance with previously published studies (e.g. Fricker et al., 2015; Wolf et al., 2012) that have attempted to use ALS data for predicting measures of tree species diversity in tropical forests. For example, Wolf et al. (2012) reported an $R^2$ of 0.48 when predicting tree species richness in the Neotropical forests of Panama. There are also deviations from other studies that have reported relatively better prediction accuracy in terms $R^2$ and RMSE (Hernández-Stefanoni et al., 2014; Laurin et al., 2014). Such deviations are not surprising, however, since in most cases the predictive ability of the remotely sensed data varies with the ecosystem and biogeographical context (Camathias et al., 2013). Furthermore, the differences in factors such as number of species, field design, data characteristics, and topography may influence the prediction accuracy. Generally, Paper IV provided an insight into how measures of tree species diversity potentially can be predicted by using ALS data in the miombo woodlands.

Table 2. Pseudo-$R^2$, absolute root mean square error (RMSE), and relative root mean square error (RMSE%) for predicted tree species richness and Shannon diversity index

| Measures of tree species diversity | Vegetation type [a] | n [b] | LMMs | | | k-NN | |
|---|---|---|---|---|---|---|---|
| | | | $R^2$ | RMSE | RMSE (%) | RMSE | RMSE (%) |
| Tree species richness | Woodlands | 387 | 0.41 | 2.08 | 38.9 | 2.84 | 53.0 |
| | Forest | 40 | 0.34 | 2.79 | 41.8 | 2.99 | 44.8 |
| | Other cover types | 57 | 0.36 | 1.84 | 57.0 | 2.93 | 90.7 |
| | All | 484 | 0.46 | 2.12 | 40.7 | 2.15 | 41.2 |
| Shannon diversity index | Woodlands | 387 | 0.32 | 0.42 | 35.0 | 0.59 | 49.2 |
| | Forest | 40 | 0.47 | 0.35 | 27.0 | 0.50 | 39.0 |
| | Other cover types | 57 | 0.39 | 0.53 | 73.0 | 0.72 | 99.0 |
| | All | 484 | 0.39 | 0.45 | 39.1 | 0.46 | 40.0 |

[a] Vegetation types according to MNRT (2011).
[b] Number of field plots.

## 6.0. Conclusions and future studies

The studies conducted within this thesis have provided information which may be useful for improving the current methods for assessment and quantification of forest resources in Tanzania. This is needed for forest management decision-making and for development of a REDD+ MRV system (Figure 2). More specifically, in Paper I models for estimating volume of miombo woodland trees, including total, merchantable stem, and branches volumes were developed. For each of these categories general and site-specific models were also constructed. The models are considered to be valuable for deriving information that can support management and utilization of miombo woodland resources by providing more accurate estimates of forest resources in terms of growing stock volumes. Volume models may also be used as an input in developing simulation tools that can be used to predict future conditions of forests under different management scenarios. Such a tool has recently been developed by Mugasha (2014) and the tree volume model with *dbh* and *ht* developed in Paper I has successfully been used to generate reliable information on volume stock.

Additionally, the volume models, when combined with expansion factors, may also be used in conjunction with NAFORMA data for estimation of forest AGB and national greenhouse gas emissions. Thus, at least two viable approaches for estimation of carbon stocks exist currently for the miombo woodlands of Tanzania, since there is already a more direct approach of using allometric AGB models developed by Mugasha et al. (2013). Although the models have been developed for miombo woodlands of Tanzania, they may potentially be used in the miombo woodlands of nearby countries, such as Malawi and Mozambique in the same context as described above. However, it is important to validate the models before use, and caution should be exercised when applying the models beyond the ranges of the data used for model development.

Papers II and III of the thesis have provided empirical evidence that ALS can potentially be used for prediction and estimation of AGB in the tropical forests and woodlands of Tanzania, despite a wide range of variation in forest structure and composition. Moreover, the findings have shown that selection of statistical modeling methods in ALS-based studies should be guided by the design used to determine the sample locations (e.g. nested/clustered designs) of the field inventory. In other words, the design of the field inventory should not be ignored when selecting the statistical method as this may have consequences for the accuracy of the models and finally for the precision of the estimates. NAFORMA data will in the future possibly be used in combination with various types of auxiliary data from remote sensing for monitoring of forest carbon. The statistical procedures described in Paper II may offer baseline information for such studies. Likewise, the experiences gained from the methodological part of this thesis may inform future decisions on the design of a REDD+ MRV system, including the modeling part of relationships between carbon stocks and auxiliary data. Furthermore, the findings in Paper III have shown that field plot size is an important survey parameter that must be considered when designing a field inventory to be used for estimation purposes in combination with auxiliary data from remote sensing. The analyses of the effects of plot size on ALS-assisted AGB estimates have clearly shown that both the model fit and the precision of the ALS-assisted AGB estimates increased with plot size. A firm and well justified recommendation for plot size in future surveys can hardly be given on the basis of this very limited study – limited in sample size (30 plots) and geographical extent (one study site), but the results do show a declining utility of increasing plot size beyond around 1200 $m^2$ in rainforest. This finding is nevertheless fairly well in line with findings from similar forests on other continents.

Lastly, the findings in Paper IV demonstrated that irrespective of the large number of tree species in the miombo woodlands of Tanzania, ALS holds potential for assessment of measures of tree species diversity. This paper was the first attempt to use remotely sensed data for predicting measures of tree species diversity in Tanzania, thus the results open an avenue for using ALS data when monitoring forest biodiversity, in particular tree species diversity.

Even though the four studies presented in this thesis have provided information that may be useful for improving the current methods for assessment and quantification of forest resources in Tanzania, there are still many research gaps that should be further addressed.

In Paper I, NLS was used for development of volume models. Compared to the commonly used OLS with log transformation this approach is considered to be more robust for development of tree allometric models (Sileshi, 2014). However, recently hierarchal Bayesian framework and mixed effects models have been preferred for development of tree allometric models (e.g. Cohen et al., 2013; Xiao et al., 2011). Future studies should therefore compare these approaches to OLS and NLS.

In Paper II, the *k*-NN was used as a non-parametric method for prediction of AGB using ALS data. Other non-parametric methods such as random forest have shown promising results in many of the previously studies (e.g. Hudak et al., 2008; Li et al., 2013; Mascaro et al., 2014). However, in this study random forest was not applied because the impact of dealing with clustered data in random forest is not well understood. With the recent development of a mixed effects approach for random forest by Hajjem et al. (2014), future studies may consider this as an alternative to improve the prediction accuracy. Furthermore, in this thesis ALS was the only source of auxiliary data from remote sensing. Fusing ALS data with other remotely sensed information, for example complementary data types such as optical and even hyperspectral (e.g. Dalponte et al., 2014) that may capture other properties of the trees and tree canopies than just 3D information, may improve estimates and could be considered in future studies. This may also help improving the quality in construction of continuous AGB maps and local small-area estimates of AGB, which may be useful for forest management planning at local scales (i.e. district to village level) and also for local REDD+ projects. This would be equivalent to products that are produced routinely with great success by ALS for forest management purposes in boreal forests (e.g. Næsset, 2014).

In Paper III, the maximum plot size of 3000 $m^2$ used to evaluate effects of plot size on the prediction accuracy and estimated precision of ALS-assisted AGB estimates leaves an open question as to whether there are any additional gains in precision beyond this size. Although, as stated above, there seems to be a declining utility of increasing plot size beyond a size of around 1200 $m^2$, future studies to confirm this finding should include plot sizes above 3000 $m^2$ to increase confidence in the results. Further studies of this important design property are clearly required also in other forest types than the rainforest. An obvious forest type to be considered in a Tanzanian context is the miombo woodlands. Likewise, other factors that may aid choice of an "optimal" plot size such as sample sizes, on-plot costs, traveling costs, and overall field inventory design should be considered. Furthermore, future studies should take terrain and topographical variables into account since for example traveling costs and on-plot costs as well as plot position accuracy may be influenced by these factors.

In Paper IV, an indirect modelling approach was used when developing predictive models for measures of tree species diversity. For future studies, use of more integrated statistical analyses, which directly relate tree species community information to ALS metrics without use of the estimators of tree species diversity, should be considered. Moreover, instead of focusing on the use of ALS alone, future studies could attempt to combine ALS with other remotely sensed data, such as hyperspectral images mentioned above. This will likely improve the prediction accuracy of the models, since the two types of sensor data will complement each other, especially when modelling multi-layered forests (e.g. Dalponte et al., 2008; Laurin et al., 2014).

Finally, we want to point out that this thesis had a purely technical approach, with the aim of providing empirical evidence of performance of ALS used for assisting in prediction of essential properties for management and carbon reporting in a broad sense. As a technology evolves from being a source of data for research purposes to a tool for operational data collection, economical aspects become essential, which also include choices between alternatives. ALS has become a great success for operational purposes in boreal forest inventory due to low acquisition costs relative to e.g. labor costs (Næsset, 2014). Higher accuracy of the produced information compared to alternative methodologies is also an important aspect. Thus, in the boreal context ALS-assisted methods have simply turned out to be more cost-effective compared to other methods (Bergseng et al., 2015; Eid et al., 2004). In developing countries in general and in Tanzania in particular, labor costs are lower and ALS acquisition costs are higher – even in absolute terms, than in many developed countries. Therefore, the economical aspects of the ALS-based applications compared to alternatives involving for example more extensive use of field campaigns, must accompany further research aiming at confirming and expanding the technical knowledge gained by this thesis -- the first comprehensive study on use of ALS in forests in Tanzania.

## References

Abbot, P., Lowore, J., Werren, M., 1997. Models for the estimation of single tree volume in four Miombo woodland types. Forest Ecology and Management 97, 25-37.

Abdallah, J., Monela, G., 2007. Overview of Miombo woodlands in Tanzania. MITMIOMBO– management of indigenous tree species for ecosystem restoration and wood production in semi-arid Miombo woodlands in eastern Africa. Working Papers of the Finnish Forest Research Institute 50, 9-23.

Angelsen, A., Brockhaus, M., 2009. Realising REDD+: National strategy and policy options. Center for International Forestry (CIFOR), Bogor, Indonesia.

Araya, M.M., Hofstad, O., 2014. Monetary incentives to avoid deforestation under the Reducing emissions from deforestation and degradation (REDD)+ climate change mitigation scheme in Tanzania. Mitigation and Adaptation Strategies for Global Change, 1-23.

Asner, G.P., Clark, J.K., Mascaro, J., Vaudry, R., Chadwick, K.D., Vieilledent, G., Rasamoelina, M., Balaji, A., Kennedy-Bowdoin, T., Maatoug, L., 2012a. Human and environmental controls over aboveground carbon storage in Madagascar. Carbon balance and management 7.1:13.

Asner, G.P., Mascaro, J., 2014. Mapping tropical forest carbon: Calibrating plot estimates to a simple LiDAR metric. Remote Sensing of Environment 140, 614-624.

Asner, G.P., Mascaro, J., Muller-Landau, H.C., Vieilledent, G., Vaudry, R., Rasamoelina, M., Hall, J.S., van Breugel, M., 2012b. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia 168, 1147-1160.

Backéus, I., Pettersson, B., Strömquist, L., Ruffo, C., 2006. Tree communities and structural dynamics in miombo (Brachystegia–Julbernardia) woodland, Tanzania. Forest Ecology and Management 230, 171-178.

Bergseng, E., Orka, H.O., Næsset, E., Gobakken, T., 2015. Assessing forest inventory information obtained from different inventory approaches and remote sensing data sources. Annals of Forest Science 72, 33-45.

Birdsey, R., Angeles-Perez, G., Kurz, W.A., Lister, A., Olguin, M., Pan, Y., Wayson, C., Wilson, B., Johnson, K., 2013. Approaches to monitoring changes in carbon stocks for REDD+. Carbon Management 4, 519-537.

Bjorndalen, J.E., 1992. Tanzania's vanishing rain forests — assessment of nature conservation values, biodiversity and importance for water catchment, in: Paoletti, M.G., Pimentel, D. (Eds.), Biotic Diversity in Agroecosystems. Elsevier, Amsterdam, pp. 313-334.

Blomley, T., Iddi, S., 2009. Participatory Forest Management in Tanzania 1993-2009: Lessons learned and experiences to date. United Republic of Tanzania, Ministry of Natural Resources and Tourism, Forestry and Beekeeping Division, p. 71.

Bollandsås, O.M., Maltamo, M., Gobakken, T., Næsset, E., 2013. Comparing parametric and non-parametric modelling of diameter distributions on independent data using airborne laser scanning in a boreal conifer forest. Forestry 86, 493-501.

Brandon, K., 2014. Ecosystem Services from Tropical Forests: Review of Current Science. CGD Working Paper 380, Center for Global Development, Washington DC.

Camathias, L., Bergamini, A., Küchler, M., Stofer, S., Baltensweiler, A., 2013. High-resolution remote sensing data improves models of species richness. Applied Vegetation Science 16, 539-551.

Campbell, B., Angelsen, A., Cunningham, A., Katerere, Y., Sitoe, A., Wunder, S., 2007. Miombo woodlands--opportunities and barriers to sustainable forest management. CIFOR, Bogor, Indonesia. http://www. cifor. cgiar. org/miombo/docs/Campbell_ Barriersand Opportunities. pdf. Accessed 23rd Aug 2015.

Chamshama, S., Mugasha, A., Zahabu, E., 2004. Stand biomass and volume estimation for Miombo woodlands at Kitulangalo, Morogoro, Tanzania. The Southern African Forestry Journal 200, 59-70.

Chiesa F., Dere M., Saltarelli, E., Sandbank, H., 2009. UN-REDD in Tanzania. Project on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries. Johns Hopkins School of Advanced International Studies, Washington DC and UNEP World Conservation Monitoring Centre, Cambridge, UK.

Cohen, R., Kaino, J., Okello, J., Bosire, J., Kairo, J., Huxham, M., Mencuccini, M., 2013. Propagating uncertainty to estimates of above-ground biomass for Kenyan mangroves: A scaling procedure from tree to landscape level. Forest Ecology and Management 310, 968-982.

Coops, N.C., Wulder, M.A., Culvenor, D.S., St-Onge, B., 2004. Comparison of forest attributes extracted from fine spatial resolution multispectral and lidar data. Canadian Journal of Remote Sensing 30, 855-866.

Dalponte, M., Bruzzone, L., Gianelle, D., 2008. Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. Geoscience and Remote Sensing, IEEE Transactions on 46, 1416-1427.

Dalponte, M., Orka, H.O., Ene, L.T., Gobakken, T., Næsset, E., 2014. Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data. Remote Sensing of Environment 140, 306-317.

De Sy, V., Herold, M., Achard, F., Asner, G.P., Held, A., Kellndorfer, J., Verbesselt, J., 2012. Synergies of multiple remote sensing data sources for REDD+ monitoring. Current Opinion in Environmental Sustainability 4, 696-706.

Dewees, P.A., Campbell, B.M., Katerere, Y., Sitoe, A., Cunningham, A.B., Angelsen, A., Wunder, S., 2010. Managing the Miombo woodlands of southern Africa: policies, incentives and options for the rural poor. Journal of Natural Resources Policy Research 2, 57-73.

Eggleston, S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K., 2006. IPCC guidelines for national greenhouse gas inventories. Institute for Global Environmental Strategies, Hayama, Japan.

Ehara, M., Hyakumura, K., Yokota, Y., 2014. REDD+ initiatives for safeguarding biodiversity and ecosystem services: harmonizing sets of standards for national application. Journal of Forest Research 19, 427-436.

Eid, T., Gobakken, T., Næsset, E., 2004. Comparing stand inventories for large areas based on photo-interpretation and laser scanning by means of cost-plus-loss analyses. Scandinavian Journal of Forest Research 19, 512-523.

FAO, 2001. Global Ecological Zoning for the Global Forest Resources Assessment 2000, Final Report. Food and Agriculture Organization of the United Nations.Working Paper, No56, Rome, p.199.

FAO, 2010a. Global forest resources assessment 2010: Main report. Food and Agriculture Organization of the United Nations. Forestry Paper, No. 163, Rome, p. 378.

FAO, 2010b. Global Forest Resources Assessment Country Report. United Republic of Tanzania. Food and Agricultural Organization of the United Nations, Rome, p. 56.

Frazer, G.W., Magnussen, S., Wulder, M.A., Niemann, K.O., 2011. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sensing of Environment 115, 636-649.

Fricker, G.A., Wolf, J.A., Saatchi, S.S., Gillespie, T.W., 2015. Predicting spatial variations of tree species richness in tropical forests from high resolution remote sensing. Ecological Applications., http://dx.doi.org/10.1890/14-1593.1.

Frost, P., 1996. The ecology of miombo woodlands. In B. Campbell (Ed.), *The Miombo in transition: Woodlands and welfare in Africa* (pp. 11–57). Bogor, Indonesia: CFIOR.

Gagliasso, D., Hummel, S., Temesgen, H., 2014. A comparison of selected parametric and non-parametric imputation methods for estimating forest biomass and basal area. Open Journal of Forestry 4, 42.

Gibbs, H.K., Brown, S., Niles, J.O., Foley, J.A., 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. Environmental Research Letters 2, 045023.

Gleason, C.J., Im, J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sensing of Environment 125, 80-91.

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Holm, S., Ørka, H.O., Astrup, R., 2012. Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment 123, 443-456.

Goetz, S., Hansen, M., Houghton, R.A., Walker, W., Laporte, N.T., Busch, J., 2015. Measurement and monitoring for REDD+: The needs, current technological capabilities, and future potential. Center for Global Development Working Paper.

GOFC-GOLD, A., 2011. A Sourcebook of Methods and Procedures for Monitoring and Reporting Anthropogenic Greenhouse Gas Emissions and Removals Caused by Deforestation, Gains and Losses of Carbon Stocks in Forests Remaining Forests, and Forestation. Natural Resources Canada, Alberta, Canada, GOFC-GOLD Report version COP17-1.

Grace, J., Mitchard, E., Gloor, E., 2014. Perturbations in the carbon budget of the tropics. Global Change Biology 20, 3238-3255.

Gregoire, T.G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., Holm, S., 2010. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research 41, 83-95.

Hajjem, A., Bellavance, F., Larocque, D., 2014. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84, 1313-1328.

Hamilton, A.C., Bensted-Smith, R., 1989. Forest conservation in the East Usambara mountains, Tanzania. IUCN, Gland, Switzerland.

Hansen, E.H., Gobakken, T., Bollandsås, O.M., Zahabu, E., Næsset, E., 2015. Modeling aboveground biomass in dense tropical submontane rainforest Using Airborne Laser Scanner Data. Remote Sensing 7, 788-807.

Hauglin, M., Gobakken, T., Lien, V., Bollandsås, O.M., Næsset, E., 2012. Estimating potential logging residues in a boreal forest by airborne laser scanning. Biomass and Bioenergy 36, 356-365.

Henry, M., Picard, N., Trotta, C., Manlay, R.J., Valentini, R., Bernoux, M., Saint-André, L., 2011. Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. Silva Fennica 45, 477-569.

Hernández-Stefanoni, J.L., Dupuy, J.M., Johnson, K.D., Birdsey, R., Tun-Dzul, F., Peduzzi, A., Caamal-Sosa, J.P., Sánchez-Santos, G., López-Merlin, D., 2014. Improving Species Diversity and Biomass Estimates of Tropical Dry Forests Using Airborne LiDAR. Remote Sensing 6, 4741-4763.

Hewson, J., Steininger, M., Pesmajoglou, S., 2014. REDD+ Measurement, Reporting and Verification (MRV) Manual, Version 2.0. USAID-supported Forest Carbon, Markets and Communities Program, Washington, DC, USA.

Hou, Z., Xu, Q., Tokola, T., 2011. Use of ALS, Airborne CIR and ALOS AVNIR-2 data for estimating tropical forest attributes in Lao PDR. ISPRS Journal of Photogrammetry and Remote Sensing 66, 776-786.

Houghton, R.A., 2013. The emissions of carbon from deforestation and degradation in the tropics: past trends and future potential. Carbon Management 4, 539-546.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., Falkowski, M.J., 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment 112, 2232-2245.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., Maltamo, M., 2008. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. International Journal of Remote Sensing 29, 1339-1366.

Kaasalainen, S., Holopainen, M., Karjalainen, M., Vastaranta, M., Kankare, V., Karila, K., Osmanoglu, B., 2015. Combining Lidar and Synthetic Aperture Radar Data to Estimate Forest Biomass: Status and Prospects. Forests 6, 252-270.

Lanly, J.-P., 2003. Deforestation and forest degradation factors, Congress Proceedings B, XII World Forestry Congress, pp. 75-83.

Laurin, G.V., Chan, J.C.-W., Chen, Q., Lindsell, J.A., Coomes, D.A., Guerriero, L., Frate, F.D., Miglietta, F., Valentini, R., 2014. Biodiversity mapping in a tropical west African forest with airborne hyperspectral data. Plos One, e97910.

Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J., 2002. Lidar Remote Sensing for Ecosystem Studies Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. BioScience 52, 19-30.

Leutner, B.F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., Wegmann, M., 2012. Modelling forest α-diversity and floristic composition—On the added value of LiDAR plus hyperspectral remote sensing. Remote Sensing 4, 2818-2845.

Lewis, S.L., 2006. Tropical forests and the changing earth system. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 195-210.

Li, J., Hu, B., Noland, T.L., 2013. Classification of tree species based on structural features derived from high density LiDAR data. Agricultural and Forest Meteorology 171-172, 104-114.

Lin, L., Sills, E., Cheshire, H., 2014. Targeting areas for Reducing Emissions from Deforestation and forest Degradation (REDD+) projects in Tanzania. Global Environmental Change 24, 277-286.

Lu, D., 2006. The potential and challenge of remote sensing-based biomass estimation. International journal of remote sensing 27, 1297-1328.

Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., Moran, E., 2014. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. International Journal of Digital Earth, 1-43.

Lucas, R.M., Mitchell, A.L., Rosenqvist, A., Proisy, C., Melius, A., Ticehurst, C., 2007. The potential of L-band SAR for quantifying mangrove characteristics and change: case studies from the tropics. Aquatic conservation: marine and freshwater ecosystems 17, 245-264.

Lupala, Z., Lusambo, L., Ngaga, Y., Makatta, A.A., 2015. The land use and cover change in miombo woodlands under community based forest management and its implication to climate change mitigation: a case of Southern Highlands of Tanzania. International Journal of Forestry Research 2015., http://dx.doi.org/10.1155/2015/459102.

Malimbwi, R., Solberg, B., Luoga, E., 1994. Estimation of biomass and volume in miombo woodland at Kitulangalo Forest Reserve, Tanzania. Journal of Tropical Forest Science 7, 230-242.

Maltamo, M., Mustonen, K., Hyyppä, J., Pitkänen, J., Yu, X., 2004. The accuracy of estimating individual tree variables with airborne laser scanning in a boreal nature reserve. Canadian Journal of Forest Research 34, 1791-1801.

Marshall, A., Willcock, S., Platts, P., Lovett, J., Balmford, A., Burgess, N., Latham, J., Munishi, P., Salter, R., Shirima, D., 2012. Measuring and modelling above-ground carbon and tree allometry along a tropical elevation gradient. Biological Conservation 154, 20-33.

Mascaro, J., Asner, G.P., Dent, D.H., DeWalt, S.J., Denslow, J.S., 2012. Scale-dependence of aboveground carbon accumulation in secondary forests of Panama: A test of the intermediate peak hypothesis. Forest Ecology and Management 276, 62-70.

Mascaro, J., Asner, G.P., Knapp, D.E., Kennedy-Bowdoin, T., Martin, R.E., Anderson, C., Higgins, M., Chadwick, K.D., 2014. A tale of two "forests": Random Forest machine learning aids tropical forest carbon mapping. PLoS one 9, e85993.

Mascaro, J., Detto, M., Asner, G.P., Muller-Landau, H.C., 2011. Evaluating uncertainty in mapping forest carbon with airborne LiDAR. Remote Sensing of Environment 115, 3770-3774.

Masota, A., Zahabu, E., Malimbwi, R., Bollandsås, O.M., Eid, T., 2015. Tree allometric models for predicting above- and belowground biomass of tropical rainforests in Tanzania. Southern Forests: a Journal of Forest Science (under revision).

Mattsson, E., Persson, U.M., Ostwald, M., Nissanka, S., 2012. REDD+ readiness implications for Sri Lanka in terms of reducing deforestation. Journal of Environmental Management 100, 29-40.

Mbwambo, L., Eid, T., Malimbwi, R., Zahabu, E., Kajembe, G., Luoga, E., 2012. Impact of decentralised forest management on forest resource conditions in Tanzania. Forests, Trees and Livelihoods 21, 97-113.

McMahon, P., 2014. A Burning Issue: Tropical Forests and the Health of Global Ecosystems. Challenges and Opportunities for the World's Forests in the 21st Century. Springer, pp. 23-35.

McRoberts, R.E., Andersen, H.-E., Næsset, E., 2014. Using Airborne Laser Scanning Data to Support Forest Sample Surveys. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Forestry Applications of Airborne Laser Scanning. Springer, pp. 269-292.

MNRT, 2011. NAFORMA Field Manual - Biophysical. http://http://www.fao.org/forestry/23484-05b4a32815ecc769685b21b03be44ea77.pdf. Accessed 23rd Feb 2015.

MNRT, 2015. National Forest Resources Monitoring and Assessment of Tanzania Mainland (NAFORMA). Main Results. http://www.fao.org/forestry/43612-09cf2f02c20b55c1c00569e679197dcde.pdf. Accessed 17th Aug 2015.

Mugasha, A.G., 2014. Decision support tools for management of miombo woodlands, Ecology and Natural Resource Management. Ph.D .Thesis, Norwergian University of Life Sciences, Norway, p. 34.

Mugasha, W.A., Eid, T., Bollandsås, O.M., Malimbwi, R.E., Chamshama, S.A.O., Zahabu, E., Katani, J.Z., 2013. Allometric models for prediction of above- and belowground biomass of trees in the miombo woodlands of Tanzania. Forest Ecology and Management 310, 87-101.

Muller, J., Moning, C., Baessler, C., Heurich, M., Brandl, R., 2009. Using airborne laser scanning to model potential abundance and assemblages of forest passerines. Basic and Applied Ecology 10, 671-681.

Munishi, P., Shear, T., 2004. Carbon storage in afromontane rain forests of the Eastern Arc Mountains of Tanzania: Their net contribution to atmospheric carbon. Journal of Tropical Forest Science 16, 78-93.

Mwakalukwa, E.E., Meilby, H., Treue, T., 2014. Volume and aboveground biomass models for dry Miombo Woodland in Tanzania. International Journal of Forestry Research 2014., http://dx.doi.org/10.1155/2014/531256.

Müller, J., Vierling, K., 2014. Assessing Biodiversity by Airborne Laser Scanning. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer, pp. 357-374.

Næsset, E., 2004. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research 19, 164-179.

Næsset, E., 2014. Area-Based Inventory in Norway–From Innovation to an Operational Reality. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer, pp. 215–240.

Næsset, E., Bollandsås, O.M., Gobakken, T., Solberg, S., McRoberts, R.E., 2015. The effects of field plot size on model-assisted estimation of aboveground biomass change using multitemporal interferometric SAR and airborne laser scanning data. Remote Sensing of Environment 168, 252-264.

Næsset, E., Gobakken, T., Solberg, S., Gregoire, T.G., Nelson, R., Ståhl, G., Weydahl, D., 2011. Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. Remote Sensing of Environment 115, 3599-3614.

Oldeland, J., Wesuls, D., Rocchini, D., Schmidt, M., Jürgens, N., 2010. Does using species abundance data improve estimates of species diversity from remotely sensed spectral heterogeneity? Ecological Indicators 10, 390-396.

Patenaude, G., Milne, R., Dawson, T.P., 2005. Synthesis of remote sensing approaches for forest carbon estimation: reporting to the Kyoto Protocol. Environmental Science & Policy 8, 161-178.

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K., Hyyppä, J., 2015. Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. Remote Sensing of Environment 158, 431-440.

Sileshi, G.W., 2014. A critical review of forest biomass estimation models, common mistakes and corrective measures. Forest Ecology and Management 329, 237-254.

Sinha, S., Jeganathan, C., Sharma, L., Nathawat, M., 2015. A review of radar remote sensing for biomass estimation. International Journal of Environmental Science and Technology, 1-14.

Swetnam, R.D., Fisher, B., Mbilinyi, B.P., Munishi, P.K.T., Willcock, S., Ricketts, T., Mwakalila, S., Balmford, A., Burgess, N.D., Marshall, A.R., Lewis, S.L., 2011. Mapping socio-economic scenarios of land cover change: A GIS method to enable ecosystem service modelling. Journal of Environmental Management 92, 563-574.

Tabor, K., Burgess, N.D., Mbilinyi, B.P., Kashaigili, J.J., Steininger, M.K., 2010. Forest and woodland cover and change in coastal Tanzania and Kenya, 1990 to 2000. Journal of East African Natural History 99, 19-45.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H.M., Chamuya, N., Zahabu, E., Otieno, J., 2014. A sampling design for a large area forest inventory: case Tanzania. Canadian Journal of Forest Research 44, 931-948.

Treue, T., Ngaga, Y., Meilby, H., Lund, J.F., Kajembe, G., Iddi, S., Blomley, T., Theilade, I., Chamshama, S., Skeie, K., 2014. Does participatory forest management promote sustainable forest utilisation in Tanzania? International Forestry Review 16, 23-38.

United Republic of Tanzania (URT), 2012. National strategy for Reduced Emissions from Deforestation and Forest Degradation (REDD+). Vice president's office, Dar es Salaam, Tanzania, p. 61.

Vanderhaegen, K., Verbist, B., Hundera, K., Muys, B., 2015. REALU vs. REDD+: Carbon and biodiversity in the Afromontane landscapes of SW Ethiopia. Forest Ecology and Management 343, 22-33.

Vauhkonen, J., Korpela, I., Maltamo, M., Tokola, T., 2010. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. Remote Sensing of Environment 114, 1263-1276.

Vauhkonen, J., Maltamo, M., McRoberts, R.E., Næsset, E., 2014. Introduction to Forestry Applications of Airborne Laser Scanning. In: Maltamo M, Næsset E,Vauhkonen J, editors. Forestry applications of airborne laser scanning –concepts and case studies. Dordrecht, Netherlands: Springer, pp. 1-16.

Venter, O., Koh, L.P., 2012. Reducing emissions from deforestation and forest degradation (REDD+): game changer or just another quick fix? Annals of the New York Academy of Sciences 1249, 137-150.

Vierling, K., Bässler, C., Brandl, R., Vierling, L., Weiß, I., Müller, J., 2011. Spinning a laser web: predicting spider distributions using LiDAR. Ecological Applications 21, 577-588.

Vogeler, J.C., Hudak, A.T., Vierling, L.A., Evans, J., Green, P., Vierling, K.T., 2014. Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. Remote Sensing of Environment 147, 13-22.

Wertz-Kanounnikoff, S., Kongphan-apirak, M., 2009. Emerging REDD+: a preliminary survey of demonstration and readiness activities. Working paper 46. Center for International Forestry Research (CIFOR). Bogor, Indonesia.

White, F., 1983. The vegetation of Africa: a descriptive memoir to accompany the UNESCO/AETFAT/UNSO vegetation map of Africa by F White. Natural Resources Research Report XX, UNESCO, Paris, France.

Wolf, J.A., Fricker, G.A., Meyer, V., Hubbell, S.P., Gillespie, T.W., Saatchi, S.S., 2012. Plant species richness is associated with canopy height and topography in a neotropical forest. Remote Sensing 4, 4010-4021.

Xiao, X., White, E.P., Hooten, M.B., Durham, S.L., 2011. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. Ecology 92, 1887-1894.

Yao, W., Krull, J., Krzystek, P., Heurich, M., 2014. Sensitivity Analysis of 3D Individual Tree Detection from LiDAR Point Clouds of Temperate Forests. Forests 5, 1122-1142.

Zahabu, E., Skutsch, M.M., Sosovele, H., Malimbwi, R.E., 2007. Reduced emissions from deforestation and degradation. African Journal of Ecology 45, 451-453.

Zolkos, S., Goetz, S., Dubayah, R., 2013. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment 128, 289-298.

# PAPER I

Southern Forests 2014: 1–11
Printed in South Africa — All rights reserved
This is the final version of the article that is
published ahead of the print and online issue


SOUTHERN FORESTS
ISSN 2070-2620 EISSN 2070-2639
http://dx.doi.org/10.2989/20702620.2014.957594

# Models for estimation of tree volume in the miombo woodlands of Tanzania

Ernest W Mauya[1]*, Wilson A Mugasha[1], Eliakimu Zahabu[2], Ole M Bollandsås[1] and Tron Eid[1]

[1] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway
[2] Department of Forest Mensuration and Management, Sokoine University of Agriculture, Morogoro, Tanzania
* Corresponding author, e-mail: ernestmauya@yahoo.com

**Volume of trees is an important parameter in forest management, but only volume models with limited geographical and tree size coverage have previously been developed for Tanzanian miombo woodlands. This study developed models for estimating total, merchantable stem and branches volume applicable for the entire miombo woodlands of Tanzania. We used data from 158 destructively sampled trees, including 55 tree species collected from wide geographical and biophysical ranges. We developed general and site-specific models with diameter at breast height only as the independent variable, together with models with both diameter at breast height and tree height. Leave-one-out cross-validation was used to evaluate the models. The total tree volume models that included diameter at breast height and tree height had appropriate predictive capabilities with relative root mean square errors ($RMSE_r$) ranging from 30.5% to 47.6%. The $RMSE_r$ for total tree volume models with diameter at breast height only ranged from 39.9% to 48.0%. The site-specific models had slightly lower $RMSE_r$ values relative to the general models. The relative mean prediction error of the general total tree volume model with diameter at breast height and tree height was lower (0.6%) than those of the previously developed models (−30.7% to 31.2%). Based on the evaluations, we recommend the general total tree models to be applied over a wide range of geographical and biophysical conditions in Tanzania.**

**Keywords:** branches, cross-validation, destructive sampling, merchantable stem

## Introduction

Estimating tree volume is important for forest management purposes such as assessment of growing stock, timber valuation, selection of forest areas for harvests, and for growth and yield studies. Volume estimation is commonly carried out using allometric models based on diameter at breast height and total tree height, and occasionally also some measures of tree form (e.g. Clutter et al. 1983). Numerous allometric tree volume models have been developed for different tree species and forest types in Europe (Zianis et al. 2005), North America (Ter-Mikaelian and Korzukhin 1997, Jenkins et al. 2004, Zhou and Hemstrom 2010) and Australia (Grierson et al. 2000; Keith et al. 2000), but relatively few models have been developed for tropical forests types in sub-Saharan Africa, as indicated in the review by Henry et al. (2011). Moreover, the reliability of forest volume estimates using the volume models that do exist for tropical forests is questionable because there are many species, tree sizes and geographic areas that are not covered by these models (Hofstad 2005; Henry et al. 2011).

The allometry of trees varies with many factors. The relationship between diameter at breast height and total tree height is closely related to environmental factors such as soil nutrients, climate, disturbance regime, successional status and topographic position, but also to tree species and several genetic factors (Feldpausch et al. 2011). Such variation is a challenge when estimating volume and calls for models calibrated for specific conditions, tree species or environmental covariates (Vieilledent et al. 2012) or to include the different sources of variation as independent variables in the model. However, due to the large variability in conditions and the large number of species in tropical forests, it would be near impossible to collect a data set that covers a sufficient number of combinations of conditions and species to make one 'universal model' or many condition- and species-specific models. Furthermore, a 'universal' model that needs a range of input variables would be very impractical to use due to the resources needed to collect necessary data. Thus, there is always a trade-off between developing relatively precise models with a narrow range of application (or models that need a lot of data input), and less precise general models applicable for wider ranges of tree species and geographical coverage. However, if possible, good and expedient grouping that reduces the within-group allometric variation is always useful in order to develop more accurate models. The grouping could, for example, be performed directly based on allometry using statistical techniques such as cluster analyses and discriminant analyses (e.g. Vanclay 1994; Akindele and LeMay 2006; Mugasha et al. 2013a). Another solution could be to develop models for broad forest types such as tropical dry forest, tropical rainforest, tropical moist forest and tropical mountain forests (e.g. Brown 2002).

Miombo woodland is part of the dry tropical forests that is characterised by the dominance of trees in the genera

*Brachystegia* and *Julbernalia*. This woodland type covers approximately 2.7 million km² in Tanzania, Congo, Angola, Zimbabwe, Zambia, Malawi and Mozambique (e.g. Frost 1996, Thomas and Packham 2007), which corresponds to about 9% of the African land area. Although there is uncertainty related to exact numbers, it is anticipated that more than 90% of the total forest and woodland in Tanzania (340 000 km²) could be defined as miombo woodland (URT 1998, Abdallah and Monela 2007). The miombo woodland provides a wide range of biophysical goods and services ranging from the provision of food, fuel wood, medicine, construction materials and some ecosystem services such as carbon and water management services (e.g. Dewees et al. 2010). In order to ensure sustainable utilisation of the forest resources in the miombo woodlands, reliable volume estimation models need to be developed to aid monitoring of the growing stock and enhance sustainable forest management.

The review by Henry et al. (2011) showed that a total of 34 general (covering several tree species and or/sites) volume models for tropical dry forests and tropical moist deciduous forests have been developed for sub-Saharan Africa. For miombo woodlands, both species-specific and general models previously have been developed in Zimbabwe and Malawi by Banks and Burrows (1966) and Abbot et al. (1997), respectively. In Tanzania, Malimbwi and Temu (1984) developed species-specific models, while Malimbwi et al. (1994) and Chamshama et al. (2004) developed general models. The applicability of the existing models in Tanzania is, however, limited by the sample size and tree size ranges of the data used for modelling. Furthermore, the models cover only limited geographical ranges of miombo woodland within the country and thus none of them can be regarded as general models covering the whole country. For example, the most commonly used models estimating total tree volume in Tanzania (Malimbwi et al. 1994, Chamshama et al. 2004) were developed using sample trees from only one site in the eastern part of the country. This is obviously a challenge for the recently established national forest inventory in Tanzania (National Forest Resources Monitoring and Assessment [NAFORMA], e.g. URT 2010), now monitoring the forests of the entire country through a systematic permanent sample plot design.

The majority of the volume models developed for sub-Saharan Africa are total tree volume models but also some tree-sectional volumes also exist (Henry et al. 2011). Although models for total tree volume are regarded as the most important and most frequently used (Zahabu 2008), the timber licensing and pricing system in Tanzania based on volume estimation (e.g. URT 2002, Milledge and Kaale 2003), requires also that tree-sectional volume models are developed. Such models will aid in obtaining accurate quantitative information on the amount of wood for specific uses, i.e. saw timber and branch fuel wood. Malimbwi et al. (1994) and Chamshama et al. (2004) developed such models for merchantable stem and branch volumes. These models, however, have the same limitations as the total tree volume models described above.

The objective of this study was to develop models for total volume, merchantable stem volume and branch volume of trees in miombo woodlands. Our data set covered wide geographical and biophysical ranges. Both general models and site-specific models were developed, both with diameter at breast height as the sole independent variable and models also including height. We also compared the performance of our general model with those of existing models by applying them on our model development data.

## Materials and methods

### *Site description*

The study sites were spatially distributed in the Manyara, Tabora, Katavi and Lindi regions in order to cover a wide range of forest conditions within the miombo woodlands of Tanzania (Figure 1). The sites are characterised by both tropical and subtropical climates with high between-site and temporal within-site variability in temperature and rainfall. Details on locations and conditions regarding the sites (hereafter labeled according to the region where they are located) are described in Table 1. The weather conditions for all sites may be divided into three distinct seasons: a hot dry season from mid-August to the end of October, a hot wet season from November to the beginning of April and a relatively cool dry season from April to mid-August. Furthermore, two rainfall regimes exist. In the southern, south-western, central and western parts of the country, including Lindi, Katavi and Tabora, the rainy season starts in mid-November and ends in mid-May. In the north and in the northern coastal zone, including Manyara, the rain is distributed over two shorter periods (October–December and March–May).

In Manyara the dominant miombo tree species include *Brachystegia microphylla*, *Brachystegia spiciformis*, *Julbernardia globiflora*, *Albizia versicolor*, *Brachystegia boehmii*, *Combretum collinum*, *Parinari curatellifolia*, *Markhamia obtusifolia*, *Senegalia nigrescens* (*Acacia nigrescens*) and *Tamarindus indica*. In Tabora the miombo woodlands contain the majority of the commercially exploitable tree species, including *Pterocarpus angolensis*, *Afzelia quanzensis*, *Dalbergia melanoxylon*, *Burkea africana*, *Pterocarpus tinctorius* and *Swartzia madagascariensis*. In Katavi the forests consists largely of wet miombo woodland and dry deciduous forest characterised by *Senegalia* (*Acacia*) spp., *Brachystegia spiciformis*, *Combretum* spp., *Commiphora* spp., *Grewia* spp., *Kigelia africana*, *Pterocarpus angolensis*, *Sterculia quinqueloba* and *Terminalia* spp. (Munishi et al. 2011). In Lindi, the forests are characterised by dry miombo, closed dense forest and riverine miombo dominated by highly valuable timber species such as *Brachystegia* and *Julbernadia* spp. and *Pterocarpus angolensis* (Malimbwi et al. 2005).

### *Data collection*

Destructive sampling is resource demanding, and in order to cope with financial limitations we had to limit data collection to some extent and implement a conscious sampling procedure. To obtain an estimate of the actual species- and size distribution to guide the selection of sample trees, we took advantage of information from systematic sample plot inventories (40 plots with radii of 15 m for each site) carried out in the same sites, but for a different purpose. Also in the present study, we systematically distributed 15-metre-radius sample plots within each site, and from these plots we selected one or two trees purposely for destructive sampling. As sampling proceeded, we evaluated the distribution of our sample trees against the target distribution from the previous inventories in order to end up with approximately the same distributions with respect to the most frequent species and tree sizes.

A total of 158 trees including 55 different tree species were sampled. Each sample tree was recorded for diameter at breast height (dbh) and height ($h$) before felling and, in addition, the species was determined. A calliper or a diameter tape (for larger trees) was used to measure dbh, whereas $h$ was measured using Suunto and Vertex hypsometers. Number of sample trees ($n$) and their ranges of dbh and $h$ for each site are shown in Table 2. With a few exceptions, the sample trees were the same as those used by Mugasha et al. (2013b) for development of biomass models (167 sample trees). The exceptions were eight trees in Lindi, for which we had no data because the procedures of data collection for volume determination were decided after starting the biomass sampling, and one tree in Manyara with inappropriate volume data.

After felling, all sample trees were cross cut into logs ranging from 1 to 2.5 m in length, measured for mid-diameter and length, and separated into merchantable stem and branches including tops. Only trees with dbh ≥ 15 cm were considered to have both merchantable stem and branches volume as defined in previous studies (e.g. Malimbwi and Temu 1984; Malimbwi et al. 1994) but also based on timber harvesting regulations of Tanzania (URT 2002). Demarcation of the merchantable stem part (suitable for timber) was done subjectively based on the length and branching patterns of the stem. The remaining

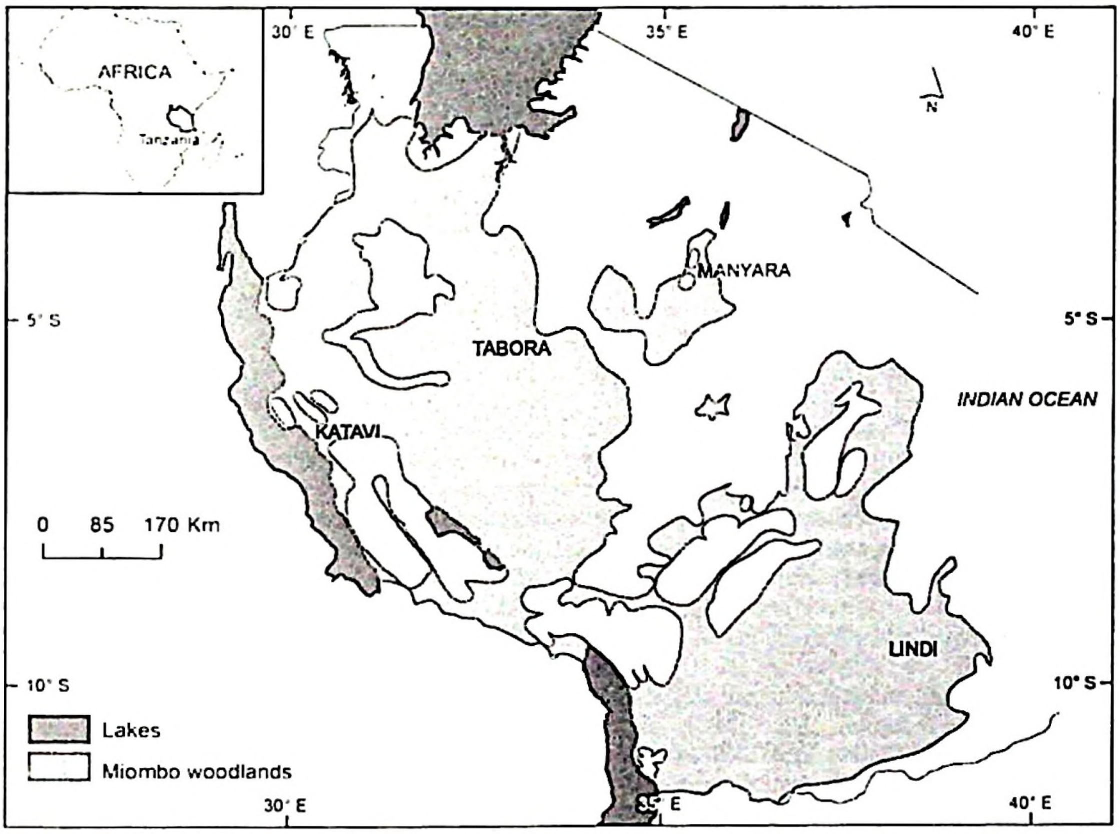



Figure 1: Study site locations and distribution of miombo woodlands in Tanzania (modified from FBD 1981)

Table 1: Study site description (source of temperature and rainfall data: Tanzania Meteorological Agency, data 1982–2012)

| Region | Forest | Location | Dominant soil type | Altitude (m) | Mean annual temperature (°C) | | Mean annual rainfall (mm) |
|---|---|---|---|---|---|---|---|
| | | | | | Min. | Max. | |
| Manyara | Ayasanda and Duru Haitemba Villages Forest Reserves | 4°20′ S, 35°47′ E | Clay alluvial soils | 1 300–1 800 | 15 | 26 | 854 |
| Lindi | Angai Villages Forest Reserve | 9°47′ S, 37°55′ E | Sandy loam soils | 330–600 | 20 | 31 | 1 915 |
| Katavi | Mpanda Ndogo general land | 6°21′ S, 30°57′ E | Sandy clay soils | 755–766 | 17 | 31 | 881 |
| Tabora | Nyahua Forest Reserve | 5°18′ S, 32°58′ E | Sandy clay loam soils | 1 096–1 103 | 17 | 30 | 771 |

tree section was considered to be branches up to a minimum diameter of 2.5 cm. Small trees with dbh < 15 cm (in total 18 trees) not suitable for timber production were not considered to have merchantable stem volume. Instead all volume was allocated to branches.

Volume of each log was calculated by multiplying the cross-sectional area at the midpoint of each log with its length. Volume of merchantable stem ($V_{stem}$) and branches ($V_{branch}$) for a tree was obtained by summing the volumes of the logs of the respective sections for that specific tree. Total tree volume ($V_{total}$) was finally obtained through summation of $V_{stem}$ and $V_{branch}$. The form factors, computed as the ratio between the measured $V_{total}$ and the volume of a cylinder having the diameter and height of the respective trees, were in average 0.76, 0.68, 0.79 and 0.76 for Manyara, Tabora, Katavi and Lindi, respectively. Table 3 shows summary statistics for merchantable stem, branches and total volume over sites and Figure 2 displays $V_{total}$ over dbh for each site.

### *Model development and evaluation*

Based on reviews of past studies on tree volume models in miombo woodlands (Hofstad 2005, Henry et al. 2011), on the general mensuration literature (e.g. Philip 1983) and on extensive initial testing, four model forms were selected and tested further. Two of the models include dbh only as the independent variable, whereas the other two include both dbh and $h$.

$$\text{Model 1: } V_x = a + b^*(\text{dbh})^2 \quad (1)$$

$$\text{Model 2: } V_x = a^*(\text{dbh})^b \quad (2)$$

$$\text{Model 3: } V_x = a^*(\text{dbh})^{b*}(h)^c \quad (3)$$

$$\text{Model 4: } V_x = a^*(\text{dbh}^2 h)^b \quad (4)$$

where $V_x$ is volume (m³) ($x$ = total, stem, or branch) and $a$, $b$ and $c$ are parameters to be estimated.

**Table 2:** Dentrometrical parameters of trees used to develop models

| Site | $n$ | dbh (cm) | | | $h$ (m) | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Min | Max | Mean | Min | Max |
| Manyara | 39 | 36.3 | 3.3 | 78.0 | 11.5 | 2.7 | 19.5 |
| Tabora | 40 | 32.1 | 1.2 | 95.0 | 12.7 | 1.9 | 26.0 |
| Katavi | 40 | 36.1 | 3.5 | 79.0 | 12.9 | 3.3 | 26.0 |
| Lindi | 39 | 30.6 | 3.5 | 76.2 | 13.0 | 3.3 | 24.0 |
| All | 158 | 33.8 | 1.2 | 95.0 | 12.5 | 1.9 | 26.0 |

Before fitting the models, the basic relationship was assessed by means of graphical plots. The graphs indicated non-linear patterns of the relationship between $V_{total}$ and dbh (Figure 2). Ordinary least square methods were applied to Model 1 whereas non-linear least square methods using the nlstools package (Baty and Delignette-Muller 2011) in R software (R Development Core Team 2013) were applied for fitting Models 2–4.

An important assumption in regression analysis is that the residual variance should be homoscedastic, or constant with increasing values of the response. This assumption is often violated in biological studies that deal with modelling of size (e.g. Parresol 1993, Özçelik et al. 2010). Thus, in the current study, to compensate for the unequal variance of the residual error when fitting the models, weighted polynomial and non-linear regression procedures using the reciprocal of dbh$^{\psi}$ as a weighting factor were applied, where $\psi$ represents the value of the weighting factor. The final value of $\psi$ was obtained by iteration, minimising the relative mean prediction error (MPE$_r$). The iteration started with an initial value of $\psi$ that was obtained using the method described by Picard et al. (2012). Let $\psi$ be 2$z$, where $z$ is the slope of the regression line. The value of $z$, and hence the initial value of $\psi$, was obtained by first dividing the independent variable (dbh) into $k$ ($k$ = 5) classes and then the variance ($\sigma_i^2$) of the dependent variable $V_x$ of each class was obtained. Then ln($\sigma_i^2$) was regressed against the natural logarithm of the mean dbh of each class, and then $z$ was obtained as the slope of the regression line as defined above.

To select the best models for $V_x$ we used leave-one-out cross-validation (e.g. James et al. 2013). In this procedure one tree at a time was excluded from the data set, while the rest of the data set (i.e. $n$ − 1) was used to calibrate a volume model by which the volume of the excluded tree was predicted. The values of $n$ for the general, site-specific, branch and merchantable data sets are presented in Table 3. From the cross-validation we calculated the root mean square error (RMSE) and the mean prediction error (MPE). Then we calculated the value of RMSE and MPE relative to the mean observed $V_x$ (RMSE$_r$ and MPE$_r$) and used those as model selection criteria. Details of their computations are:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)^2}{n}} \quad (5)$$

$$\text{RMSE}_r = \frac{\text{RMSE}}{\bar{v}_i} \times 100 \quad (6)$$

**Table 3:** Mean, minimum (Min.) and maximum (Max.) volume (m³) for merchantable stem ($V_{stem}$), branches ($V_{branch}$) and total ($V_{total}$) over sites

| Site | $V_{stem}$ | | | | $V_{branch}$ | | | | $V_{total}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | Mean | Min. | Max. | $n$ | Mean | Min. | Max. | $n$ | Mean | Min. | Max. |
| Manyara | 33 | 0.378 | 0.049 | 1.577 | 39 | 1.261 | 0.002 | 7.038 | 39 | 1.580 | 0.002 | 7.872 |
| Tabora | 31 | 0.571 | 0.281 | 2.687 | 40 | 1.003 | 0.000 | 5.189 | 40 | 1.445 | 0.000 | 7.531 |
| Katavi | 34 | 0.670 | 0.035 | 3.080 | 40 | 1.233 | 0.002 | 7.517 | 40 | 1.802 | 0.002 | 10.597 |
| Lindi | 32 | 0.384 | 0.040 | 1.849 | 39 | 0.893 | 0.002 | 5.794 | 39 | 1.208 | 0.002 | 7.594 |
| All | 130 | 0.501 | 0.028 | 3.080 | 158 | 1.097 | 0.000 | 7.517 | 158 | 1.511 | 0.000 | 10.597 |

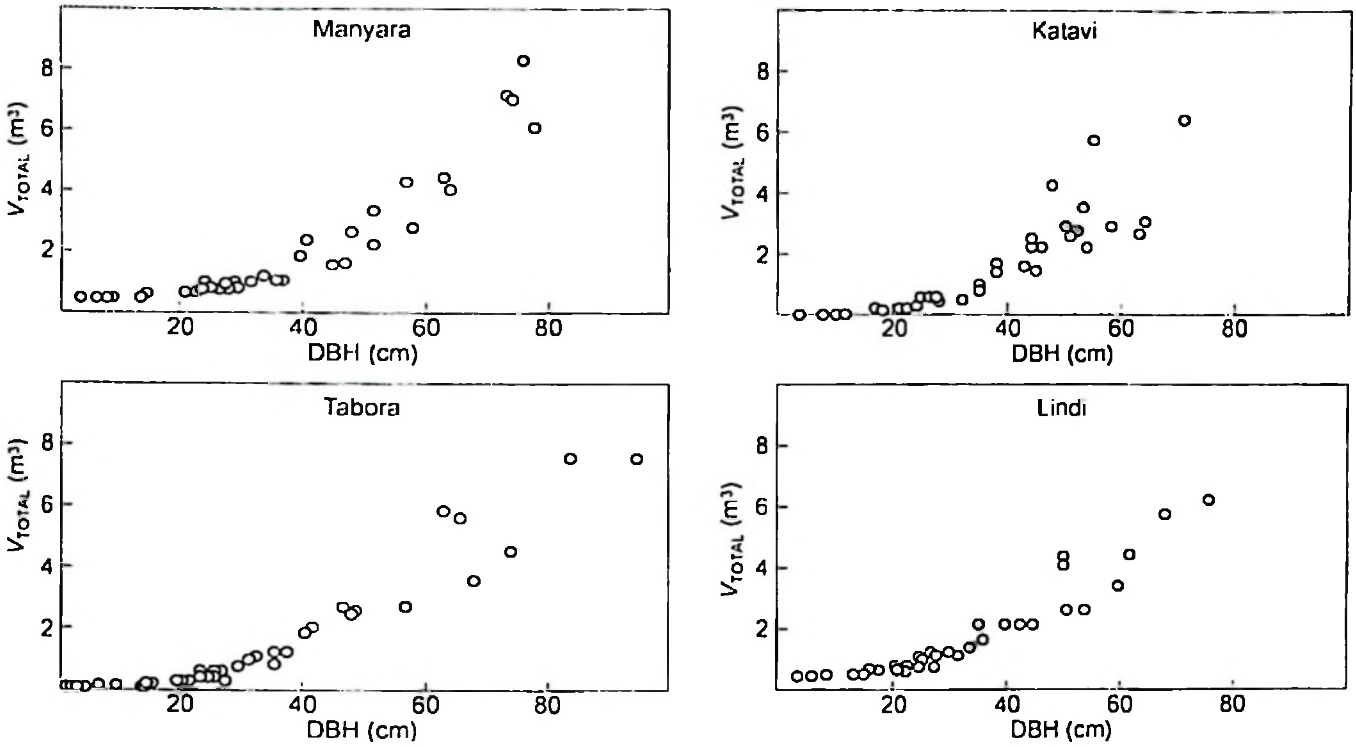



**Figure 2:** Total tree volume ($V_{total}$, m³) distribution over diameter at breast height (dbh; cm) for the four study sites

$$\mathrm{MPE} = \sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)}{n} \quad (7)$$

$$\mathrm{MPE}_r = \frac{\mathrm{MPE}}{\bar{v}_x} \times 100 \quad (8)$$

where $n$ is the number of sample trees, $V_i$ is the observed volume of tree $i$, $\hat{V}_i$ is the predicted volume of tree $i$, and $\bar{V}_x$ is the mean observed volume of the response in question ($x$ = total, stem or branch).

Paired $t$-tests between the observed and predicted volume were employed for testing the statistical significance of the MPE (bias). We also computed pseudo-$R^2$ values based on sum of squared residuals (SSR) and the corrected total sum of squares (CSST)

$$\text{pseudo-}R^2 = 1 - \left(\frac{\mathrm{SSR}}{\mathrm{CSST}}\right) \quad (9)$$

In order to reflect stability of the selected general $V_{total}$ models applied under different conditions, we partitioned the prediction values obtained from the leave-one-out cross-validation into different tree sizes (i.e. dbh and $h$ classes) and sites, and assessed the pattern and magnitude of $\mathrm{MPE}_r$. In addition, we also computed the standard deviations of residuals (STD) and relative standard deviations ($\mathrm{STD}_r$) to examine the uncertainty of the models when applied over different tree sizes and sites:

$$\mathrm{STD} = \sqrt{\sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)^2}{n-1}} \quad (10)$$

$$\mathrm{STD}_r = \left(\frac{\mathrm{STD}}{\bar{v}}\right) \times 100 \quad (11)$$

Some previously developed models were also tested on our data. Malimbwi et al. (1994) developed $V_{total}$ models based on 17 sample trees from miombo woodland in the Morogoro region, eastern Tanzania. The dbh for the sample trees ranged from 4 to 43 cm, whereas the $h$ ranged from 4 to 22 m. Two models were recommended, one with dbh only as the independent variable and the other model with both dbh and $h$ as independent variables. Chamshama et al. (2004) developed models for $V_{total}$ with both dbh and $h$ as independent variables based on data from the same site as Malimbwi et al. (1994). The models were developed from 30 sample trees with dbh ranging from 1 to 50 cm and heights ranging from 2 to 18 m. Abbot et al. (1997) developed a model for $V_{total}$ with dbh as the only independent variable based on data from four different sites of *Julbernardia–Brachystegia* woodlands in Malawi. Their cross-site model for total volume was developed from 88 sample trees with a dbh range of 5 to 35 cm.

The general models and the previously developed models were also examined by graphical plots that displayed the behaviour of each model over a wide range of diameters (also outside their respective model data ranges). Among these models, those developed by Malimbwi et al. (1994) and Chamshama et al. (2004) included $h$ as one of the independent variables. Thus, in order to account for the variation of height with diameter, we first applied a diameter–height model developed by Mugasha et al. (2013a) in order to obtain values for $h$.

The values were then used in the respective models for displaying $V_{total}$.

## Results

The parameter estimates, pseudo-$R^2$ values, observed mean volumes, RMSE$_r$, and MPE$_r$ are presented in Table 4 for the $V_{total}$ models with dbh only as the independent variable and in Table 5 for the $V_{total}$ models that include both dbh and $h$. Both model types accounted for a high proportion of the variability in the observed $V_{total}$ (high pseudo-$R^2$). For all models, the differences in pseudo-$R^2$ between dbh only, and dbh and $h$ models, were moderate. For all models, parameter estimates were significantly different from zero ($p < 0.05$).

Among the models with dbh as the only independent variable (Table 4), Model 2 was consistently the best, judged by RMSE$_r$ and MPE$_r$ values. Based on the same criteria, Model 3 was judged as the best general model in the category of models with both dbh and $h$ as independent variables (Table 5). Among the site-specific models, there were only small differences in the performance between Model 3 and Model 4, but Model 4 performed slightly better in terms of RMSE$_r$. None of the selected models (Tables 4 and 5) had MPE$_r$ values statistically significant different from zero ($p > 0.05$).

Table 6 presents model parameters and fit statistics for the selected tree section $V_{stem}$ and $V_{branch}$ models. For the general models, the results indicated that the $V_{stem}$ model performed slightly better than the $V_{branch}$ model (RMSE$_r$ 56.1% and 56.5%, respectively). The development of

**Table 4:** Fit statistics and performance criteria for total tree volume ($V_{total}$) models with dbh as the independent variable. The selected model is highlighted in bold

| Model category | $n$ | Observed volume (m³) | Model | $a$ | $b$ | pseudo-$R^2$ | RMSE$_r$ (%) | MPE$_r$ (%) |
|---|---|---|---|---|---|---|---|---|
| General | 158 | 1.511 | 1 | −0.00494 | 0.00090 | 0.86 | 49.6 | 8.1 |
| | | | **2** | **0.00016** | **2.46300** | **0.87** | **48.0** | **−0.5** |
| Manyara | 39 | 1.580 | 1 | −0.01296 | 0.00079 | 0.83 | 55.3 | 14.4 |
| | | | **2** | **0.00001** | **2.62300** | **0.95** | **32.7** | **0.6** |
| Tabora | 40 | 1.445 | 1 | −0.00155 | 0.00030 | 0.89 | 48.9 | 16.6 |
| | | | **2** | **0.00042** | **2.19786** | **0.92** | **48.0** | **−1.7** |
| Katavi | 40 | 1.802 | 1 | −0.01851 | 0.00096 | 0.77 | 58.8 | 12.4 |
| | | | **2** | **0.00009** | **2.64200** | **0.86** | **47.5** | **0.3** |
| Lindi | 39 | 1.208 | 1 | −0.01261 | 0.00086 | 0.90 | 41.1 | 10.7 |
| | | | **2** | **0.00016** | **2.47200** | **0.91** | **39.9** | **−2.3** |

**Table 5:** Fit statistics and performance criteria for total tree volume ($V_{total}$) models with *dbh* and *h* as independent variables. The selected model is highlighted in bold

| Model category | $n$ | Observed volume (m³) | Model | $a$ | $b$ | $c$ | pseudo-$R^2$ | RMSE$_r$ (%) | MPE$_r$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| General | 158 | 1.511 | **3** | **0.00011** | **2.13300** | **0.57580** | **0.88** | **47.6** | **−0.6** |
| | | | 4 | 0.00010 | 0.95530 | | 0.88 | 48.3 | 0.7 |
| Manyara | 39 | 1.580 | 3 | 0.00006 | 2.23800 | 0.66230 | 0.96 | 31.1 | 0.1 |
| | | | **4** | **0.00005** | **1.01300** | | **0.96** | **30.5** | **1.1** |
| Tabora | 40 | 1.445 | 3 | 0.00023 | 1.95535 | 0.51207 | 0.92 | 49.9 | −1.9 |
| | | | **4** | **0.00032** | **0.82890** | | **0.93** | **46.8** | **−2.0** |
| Katavi | 40 | 1.802 | 3 | 0.00006 | 2.12400 | 0.86550 | 0.93 | 34.1 | −1.1 |
| | | | **4** | **0.00006** | **1.01200** | | **0.93** | **33.3** | **−0.2** |
| Lindi | 39 | 1.208 | 3 | 0.00013 | 2.19000 | 0.49800 | 0.93 | 39.1 | −1.2 |
| | | | **4** | **0.00010** | **0.94160** | | **0.92** | **38.7** | **0.2** |

**Table 6:** Fit statistics and performance criteria for merchantable stem ($V_{stem}$) and branch ($V_{branch}$) volume models

| Model category | Tree section | $n$ | Observed volume (m³) | Model | $a$ | $b$ | pseudo-$R^2$ | RMSE$_r$ (%) | MPE$_r$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| General | $V_{stem}$ | 130 | 0.501 | 2 | 0.00011 | 2.20500 | 0.76 | 56.1 | −0.2 |
| | $V_{branch}$ | 158 | 1.097 | 2 | 0.00012 | 2.44400 | 0.85 | 56.5 | −0.4 |
| Manyara | $V_{stem}$ | 33 | 0.378 | 2 | 0.00005 | 2.35300 | 0.88 | 41.9 | 1.6 |
| | $V_{branch}$ | 39 | 1.261 | 2 | 0.00005 | 2.66700 | 0.90 | 47.6 | 0.1 |
| Tabora | $V_{stem}$ | 31 | 0.571 | 2 | 0.00033 | 1.96585 | 0.87 | 44.0 | −2.0 |
| | $V_{branch}$ | 40 | 1.003 | 2 | 0.00001 | 2.45800 | 0.86 | 56.2 | 2.2 |
| Katavi | $V_{stem}$ | 34 | 0.670 | 2 | 0.00005 | 2.47800 | 0.76 | 51.5 | 0.0 |
| | $V_{branch}$ | 40 | 1.233 | 2 | 0.00005 | 2.64200 | 0.83 | 55.5 | 0.5 |
| Lindi | $V_{stem}$ | 32 | 0.384 | 2 | 0.00015 | 2.13600 | 0.84 | 45.4 | 1.2 |
| | $V_{branch}$ | 39 | 0.893 | 2 | 0.00020 | 2.35300 | 0.90 | 46.3 | −1.0 |

site-specific models seemed to have greatest effect for $V_{stem}$ as judged by a relatively large decrease in RMSE, values for this model going from general to site-specific. MPE, values were not statistically significant different from zero ($p > 0.05$) for any of the models.

The results regarding the performance of the general $V_{total}$ models when applied over different tree sizes and sites are presented in Table 7. For the model with dbh only as independent variable, no MPE value was significantly different from zero ($p < 0.05$) in the different dbh classes. However, for low trees (the two lowest $h$ classes) there seemed to be a bias. For the model with dbh and $h$ as independent variables, the results were better with a sole significant MPE value for the smallest $h$ class. Distribution of the prediction errors over different classes of diameter-height ratio (dbh/$h$) also showed that the model with dbh and $h$ performed slightly better than the model that depends on dbh only as judged by the significance of the MPE values. None of the models seemed to produce any particular pattern of the prediction errors across the dbh and $h$ classes.

Evaluation of the prediction error of the general model with dbh only over sites showed only non-significant MPE values. However, the model with both dbh and $h$ produced a slightly significant MPE value in Katavi. These patterns can also be seen in Figure 3 where the general model with dbh and $h$ and site-specific models with dbh and $h$ are displayed for each site. The STD, values for both general models were slightly less than 50% when considering all observations (Table 7).

Table 7: Testing the selected general total tree volume ($V_{total}$) models with dbh only ($0.00016dbh^{2.45300}$), and with dbh and $h$ ($0.00011dbh^{2.13300}h^{0.57560}$) over tree different tree sizes (i.e. dbh and $h$ classes), ratio and sites

| Model category | Size class | $n$ | Observed volume ($m^3$) | MPE ($m^3$) | MPE, (%) | $p$-value for MPE | STD ($m^3$) | STD, (%) |
|---|---|---|---|---|---|---|---|---|
| General with dbh only | dbh (cm) | | | | | | | |
| | 0–20 | 40 | 0.062 | -0.007 | -11.3 | 0.15580 | 0.030 | 48.4 |
| | 20–30 | 40 | 0.414 | -0.031 | -7.5 | 0.12290 | 0.130 | 31.4 |
| | 30–48 | 39 | 1.386 | 0.086 | 6.2 | 0.28260 | 0.498 | 35.9 |
| | >50 | 39 | 4.246 | -0.075 | -1.8 | 0.73350 | 1.376 | 32.4 |
| | $h$ (m) | | | | | | | |
| | 0–8 | 40 | 0.108 | -0.040 | -37.0 | 0.00578 | 0.097 | 89.8 |
| | 8–12 | 40 | 0.596 | -0.126 | -21.1 | 0.02454 | 0.366 | 61.4 |
| | 12–16 | 39 | 1.455 | -0.037 | -2.5 | 0.96410 | 0.463 | 31.8 |
| | >16 | 39 | 3.941 | 0.146 | 3.7 | 0.49740 | 1.294 | 32.8 |
| | dbh/$h$ | | | | | | | |
| | 0.6–1.9 | 40 | 0.162 | 0.006 | 3.7 | 0.68960 | 0.093 | 57.4 |
| | 1.9–2.5 | 40 | 0.738 | 0.059 | 8.0 | 0.20980 | 0.299 | 40.5 |
| | 2.5–3.2 | 39 | 2.041 | 0.276 | 13.5 | 0.02536 | 0.792 | 38.8 |
| | >3.2 | 39 | 3.155 | -0.372 | -11.8 | 0.04773 | 1.196 | 37.9 |
| | Site | | | | | | | |
| | Manyara | 39 | 1.580 | -0.120 | -7.6 | 0.12930 | 0.920 | 58.2 |
| | Tabora | 40 | 1.445 | -0.157 | -10.9 | 0.28100 | 0.963 | 66.6 |
| | Katavi | 40 | 1.802 | 0.209 | 11.6 | 0.14600 | 0.914 | 50.7 |
| | Lindi | 39 | 1.208 | 0.038 | 3.1 | 0.57360 | 0.426 | 35.3 |
| | All | 158 | 1.511 | -0.007 | -0.5 | 0.90090 | 0.723 | 47.8 |
| General with dbh and $h$ | dbh (cm) | | | | | | | |
| | 0–20 | 40 | 0.062 | -0.008 | -12.9 | 0.10200 | 0.031 | 50.0 |
| | 20–30 | 40 | 0.414 | -0.035 | -8.5 | 0.06643 | 0.122 | 29.5 |
| | 30–48 | 39 | 1.386 | 0.079 | 5.7 | 0.27570 | 0.456 | 32.9 |
| | >50 | 39 | 4.246 | -0.073 | -1.7 | 0.74430 | 1.388 | 32.7 |
| | $h$ (m) | | | | | | | |
| | 0–8 | 40 | 0.108 | -0.012 | -11.1 | 0.04473 | 0.040 | 37.0 |
| | 8–12 | 40 | 0.596 | -0.044 | -7.4 | 0.15410 | 0.197 | 33.1 |
| | 12–16 | 39 | 1.455 | 0.036 | 2.5 | 0.60890 | 0.445 | 30.6 |
| | >16 | 39 | 3.941 | -0.016 | -0.4 | 0.94180 | 1.382 | 35.1 |
| | dbh/$h$ | | | | | | | |
| | 0.6–1.9 | 40 | 0.162 | -0.022 | -13.6 | 0.12170 | 0.091 | 56.2 |
| | 1.9–2.5 | 40 | 0.738 | -0.002 | -0.3 | 0.94980 | 0.251 | 34.0 |
| | 2.5–3.2 | 39 | 2.041 | 0.202 | 9.9 | 0.05964 | 0.681 | 33.4 |
| | >3.2 | 39 | 3.155 | -0.214 | -6.8 | 0.29170 | 1.270 | 40.3 |
| | Site | | | | | | | |
| | Manyara | 39 | 1.580 | -0.024 | -1.5 | 0.84720 | 0.470 | 29.7 |
| | Tabora | 40 | 1.445 | -0.254 | -17.6 | 0.12500 | 1.057 | 73.1 |
| | Katavi | 40 | 1.802 | 0.236 | 13.1 | 0.04684 | 0.766 | 42.5 |
| | Lindi | 39 | 1.208 | 0.005 | 0.4 | 0.94260 | 0.424 | 35.1 |
| | All | 158 | 1.511 | -0.009 | -0.6 | 0.87100 | 0.721 | 47.7 |

The results regarding the performance of the previously developed $V_{total}$ models over sites are presented in Table 8. The model with dbh only as the independent variable developed by Malimbwi et al. (1994) significantly under-predicted volume by 8.5%. In addition, the models with dbh and $h$ as independent variables developed by Malimbwi et al. (1994) and Chamshama et al. (2004) significantly under-predicted volume by 20.8% and 31.2%, respectively. The model developed by Abbot et al. (1997) significantly over-predicted the volume by 30.6%. The trends of over- and under-prediction are also seen in Figure 4 in which both the general model with both dbh and $h$ as

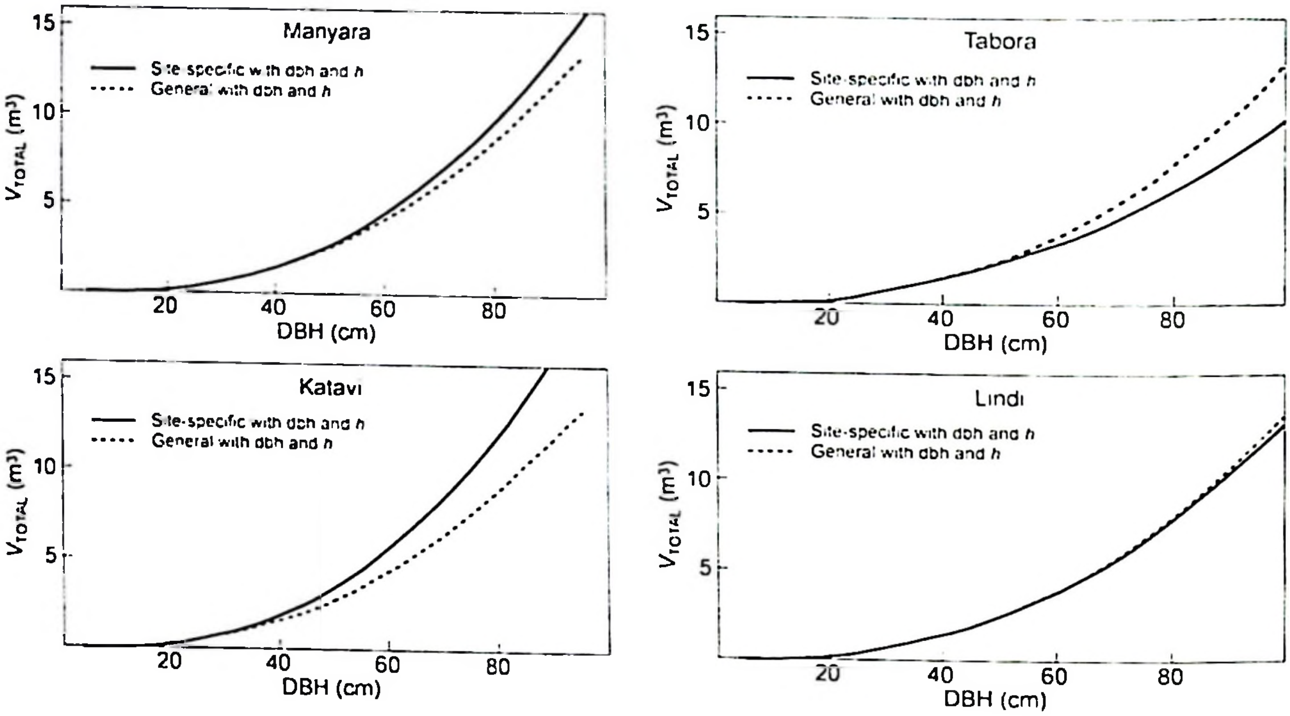



**Figure 3:** Display of the predicted total tree volume ($V_{total}$) over diameter at breast height (dbh) for the general and site-specific models with dbh and $h$ as independent variables

**Table 8:** Testing of some previously developed total tree volume ($V_{total}$) models over sites

| Model category | Site | $n$ | Observed volume (m³) | MPE (m³) | $p$-value for MPE | MPE, (%) | STD (m³) | STD, (%) |
|---|---|---|---|---|---|---|---|---|
| Malimbwi et al. (1994) with dbh only | Manyara | 39 | 1.580 | 0.027 | 0.72601 | 1.7 | 0.480 | 30.4 |
| | Tabora | 40 | 1.445 | -0.017 | 0.89104 | -1.2 | 0.794 | 55.0 |
| | Katavi | 40 | 1.802 | 0.351 | 0.01843 | 19.5 | 0.903 | 50.1 |
| | Lindi | 39 | 1.208 | 0.154 | 0.02859 | 12.8 | 0.423 | 35.1 |
| | All | 158 | 1.511 | 0.129 | 0.02014 | 8.6 | 0.692 | 45.8 |
| Malimbwi et al. (1994) with dbh and $h$ | Manyara | 39 | 1.580 | 0.326 | 0.00645 | 20.6 | 0.707 | 44.7 |
| | Tabora | 40 | 1.445 | 0.121 | 0.21297 | 8.3 | 0.603 | 41.7 |
| | Katavi | 40 | 1.802 | 0.561 | 0.00052 | 31.1 | 0.938 | 52.0 |
| | Lindi | 39 | 1.208 | 0.249 | 0.00145 | 20.6 | 0.453 | 37.5 |
| | All | 158 | 1.511 | 0.315 | 0.00000 | 20.8 | 0.711 | 47.1 |
| Chamshama et al. (2004) with dbh and $h$ | Manyara | 39 | 1.580 | 0.610 | 0.00044 | 38.6 | 0.990 | 62.6 |
| | Tabora | 40 | 1.445 | 0.197 | 0.04771 | 13.6 | 0.610 | 42.2 |
| | Katavi | 40 | 1.802 | 0.748 | 0.00003 | 41.5 | 0.997 | 55.3 |
| | Lindi | 39 | 1.208 | 0.331 | 0.00088 | 27.4 | 0.572 | 47.4 |
| | All | 158 | 1.511 | 0.472 | 0.00000 | 31.2 | 0.839 | 55.6 |
| Abbot et al. (1997) with dbh only | Manyara | 39 | 1.580 | -0.650 | 0.00100 | -41.1 | 1.139 | 72.0 |
| | Tabora | 40 | 1.445 | -0.745 | 0.04846 | -51.5 | 2.313 | 160.0 |
| | Katavi | 40 | 1.802 | -0.222 | 0.13688 | -12.3 | 0.925 | 51.3 |
| | Lindi | 39 | 1.208 | -0.234 | 0.12069 | -19.4 | 0.921 | 76.3 |
| | All | 158 | 1.511 | -0.463 | 0.00010 | -30.7 | 1.455 | 96.3 |

independent variable and the previously developed models are displayed over dbh

## Discussion

The data used in this study provided a good basis for developing versatile models for estimating volume of miombo woodland trees, both in terms of number of trees and geographical range. Within the geographical range of the study, there are different conditions in terms of climate, soil and productivity (Table 1), all factors that affect allometry and hence the relationship between the independent variables and volume. The data also covers a large range of tree sizes, which reduces the probability of having to extrapolate the model for large-size trees (see also Figure 4). For mature forests, this is of particular importance because large trees of course account for the largest part of the volume (e.g. Zahabu 2008). Although the number of trees species represented (55) is larger than in previous studies, it is far from the total number of species found in miombo woodland. Mugasha et al. (2013a), for example, found 188 different tree species in miombo woodland in a large data set including more than 30 sites covering most of Tanzania. On the other hand, our sampling was guided by the observed species distribution from an independent inventory so that we cover the most frequently occurring genera, such as *Brachystegia*, *Pterocarpus*, *Julbernadia* and *Combretum*. Such considerations were not made in the development of the models that presently are applied for volume estimation in the miombo woodland of Tanzania, i.e. the models developed by Malimbwi et al. (1994) and by Chamshama et al. (2004).

The cross-validation showed that no MPE values were significantly different from zero indicating an appropriate model performance for the selected model for $V_{total}$ (Tables 4 and 5), $V_{stem}$ (Table 6) and $V_{branch}$ (Table 6). Although the site-specific models in general were better fitted to the data compared to the general models, as would be expected (cf. Abbot et al. 1997, Jenkins et al. 2003; Djomo et al. 2010, Adekunle et al. 2013, Mugasha et al. 2013b) because allometry varies between sites, the

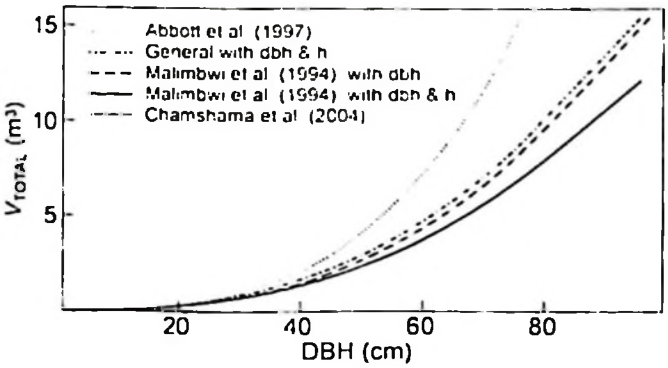



**Figure 4:** Display of predicted $V_{total}$ over dbh based on the general model with dbh and $h$ as independent variables developed in this study, and on the models developed by Abbot et al. (1997), Malimbwi et al. (1994), and Chamshama et al. (2004)

improvements were moderate. Furthermore, there were also improvements in the performance criteria (pseudo-$R^2$ and RMSE,) when $h$ was included as an independent variable (Tables 4 and 5). The improvements were moderate, however, which conform with several previous studies (e.g. Malimbwi et al. 1994, Abbot et al. 1997, Chamshama et al. 2004, Guendehou et al. 2012). Due to the generally large variation in branching patterns among tree sizes and species in miombo woodland, the fit of the $V_{branch}$ models were not as good as the fit of the $V_{total}$ and $V_{stem}$ models. It can also be noted that even though the $V_{total}$ models are affected by the branches, the model fits were still better than those of the $V_{stem}$ models. The reason for this is that the demarcation point for merchantable stem relies on considerations not only on size (minimum diameter), but also on subjective quality assessments, which adds variability to the relationship between dbh and $V_{stem}$.

Distributed over classes of tree sizes (dbh and $h$ classes), allometry (dbh/$h$) and sites (Table 7), a few MPE values significantly different from zero ($p < 0.05$) were observed. However, no particular patterns were observed except that the MPE, had a tendency to be larger for small trees. Dealing with modeling of size, this is common as absolute differences easily become large in relative terms when the observed mean is small. Furthermore, the general models seemed to give unbiased results when evaluated over sites, except for the site in Katavi where MPE of the model including both dbh and $h$ were slightly significantly different from zero. Thus, this indicates that the general models can be applied for all miombo woodlands in Tanzania with an appropriate accuracy in predictions. However, since the site-specific models perform best for their respective sites, the site-specific models should be applied in local inventories in their respective sites (see also Abbot et al. 1997, Mugasha et al. 2013b). Since the deviations between the general and site-specific models are increasing with increasing dbh (Figure 3), this is of particular importance when tree sizes are large.

Even if our data material had a wide range in terms of tree size, extrapolations beyond this range (dbh > 95 cm) sometimes have to be done when the models are applied to extremely large trees, for example in national forest inventories such as the NAFORMA. In these situations models with both dbh and $h$ as explanatory variables are preferable because $h$ moderates the effect of diameter in the model. A tree reaches maximum height much earlier than it reaches maximum diameter, so beyond the point that a tree reaches its maximum height, the increase in volume per unit increase of diameter is reduced. Models that depend on dbh only such as the ones presented here are not able to reflect this behaviour of the relationship between dbh and volume for large trees, whereas a model including $h$ to a greater extent is. If $h$ is not measured it is still possible to apply models that need $h$ as input by estimating $h$ with a height–diameter model (Mugasha et al. 2013a). However, measurements and estimations of $h$ might be associated with considerable uncertainty. Trees in miombo woodlands tend to have rounded crowns, so height measurements are prone to have random and/or systematic errors. Other authors have therefore recommended the use of models with dbh as sole predictor (e.g. Overman et al. 1994; Segura and Kanninen 2005). Nevertheless, our results suggest that models including both dbh and *h* perform equally well within the ranges of the data material.

The importance of being cautious when applying models beyond both size- and geographical ranges is illustrated by the results from the tests of previously developed $V_{tota}$ models on our data (Table 8). All tested models (Malimbwi et al. 1994, Abbot et al. 1997, Chamshama et al. 2004) were developed from data with a more narrow range of tree sizes. In addition, the model of Abbot et al. (1997) was based on data from a totally different geographic location (Malawi).

## Conclusions

We have developed models for estimating volume of miombo woodland trees including total, merchantable stem and branches volumes. Models that use dbh and *h* as independent variables were most suitable for tree volume estimation. Alternative models that utilise dbh alone were also developed. Evaluations revealed that the general total tree models can be applied over a wide range of geographical and biophysical conditions in Tanzania with an appropriate accuracy in predictions. However, we recommend that the site-specific models should be applied for local inventories in their respective sites.

*Acknowledgements* — The work has been done as part of the project 'Development of biomass estimation models and carbon monitoring in selected vegetation types of Tanzania' under the Climate Change Impacts, Adaptation and Mitigation programme (CCIAM) in Tanzania. Financial support for field work has also been provided by the Norwegian University of Life Sciences. Thanks also to district forest staff in the Liwale, Sikonge, Babati and Mpanda districts for their support during the field work. We thank the Forest and Beekeeping division of the Ministry of Natural Resources and Tourism (MNRT) in Tanzania for giving us a permit to conduct destructive sampling in all the study sites.

## References

Abbot P, Lowore J, Werren M. 1997. Models for the estimation of single tree volume in four Miombo woodland types. *Forest Ecology and Management* 97: 25–37.

Abdallah J, Monela G. 2007. Overview of Miombo woodlands in Tanzania. MITMIOMBO–management of indigenous tree species for ecosystem restoration and wood production in semi-arid Miombo woodlands in eastern Africa. *Working Papers of the Finnish Forest Research Institute* 50: 9–23.

Adekunle V, Nair K, Srivastava A, Singh N. 2013. Models and form factors for stand volume estimation in natural forest ecosystems: a case study of Katarniaghat Wildlife Sanctuary (KGWS), Bahraich District, India. *Journal of Forestry Research* 24: 217–226.

Akindele S, LeMay V. 2006. Development of tree volume equations for common timber species in the tropical rain forest area of Nigeria. *Forest Ecology and Management* 226: 41–48.

Banks P, Burrows P. 1966. Preliminary local volume tables for *Baikiaea plurijuga*, *Guibourtia coleosperma* and *Pterocarpus angolensis* in Rhodesia. *Commonwealth Forest Review* 45: 256–264.

Baty F, Delignette-Muller M-L. 2011. nlstools: tools for nonlinear regression diagnostics. R package. Available at http://cran.r-project.org/web/packages/nlstools/index.html [accessed 6 June 2013].

Brown S. 2002. Measuring carbon in forests: current status and future challenges. *Environmental Pollution* 116: 363–372.

Chamshama S, Mugasha A, Zahabu E. 2004. Stand biomass and volume estimation for miombo woodlands at Kitulangalo, Morogoro, Tanzania. *Southern African Forestry Journal* 200: 59–70.

Clutter JL, Fortson JC, Pienaar LV, Brister GH, Bailey RL. 1983. *Timber management: a quantitative approach*, vol. 6. New York: Wiley.

Dewees PA, Campbell BM, Katerere Y, Sitoe A, Cunningham AB, Angelsen A, Wunder S. 2010. Managing the Miombo woodlands of southern Africa: policies, incentives and options for the rural poor. *Journal of Natural Resources Policy Research* 2: 57–73.

Djomo AN, Ibrahima A, Saborowski J, Gravenhorst G. 2010. Allometric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. *Forest Ecology and Management* 260: 1873–1885.

FBD (Forestry and Beekeeping Division). 1981. Vegetation distribution and potential areas for beekeeping in Tanzania. Map no. JB 2054.

Feldpausch T, Banin L, Phillips O, Baker T, Lewis S, Quesada C, Affum-Baffoe K, Arets E, Berry N, Bird M. 2011. Height-diameter allometry of tropical forest trees. *Biogeosciences* 8: 1081–1106.

Frost P. 1996. The ecology of miombo woodlands. In: Campbell BM (ed.), *The miombo in transition: woodlands and welfare in Africa*. Bogor: Center for International Forestry Research. pp 11–57.

Grierson PF, Adams MA, Williams K. 2000. Review of unpublished biomass-related information: Western Australia, South Australia, New South Wales and Queensland. Canberra: Australian Greenhouse Office.

Guendehou G, Lehtonen A, Moudachirou M, Mäkipää R, Sinsin B. 2012. Stem biomass and volume models of selected tropical tree species in West Africa. *Southern Forests* 74: 77–88.

Henry M, Picard N, Trotta C, Manlay RJ, Valentini R, Bernoux M, Saint-André L. 2011. Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. *Silva Fennica* 45: 477–569.

Hofstad O. 2005. Review of biomass and volume functions for individual trees and shrubs in southeast Africa. *Journal of Tropical Forest Sciences* 17: 151–62.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning*. New York: Springer-Verlag.

Jenkins JC, Chojnacky DC, Heath LS, Birdsey RA. 2003. National-scale biomass estimators for United States tree species. *Forest Science* 49: 12–35.

Jenkins JC, Chojnacky DC, Heath LS, Birdsey RA. 2004. Comprehensive database of diameter-based biomass regressions for North American tree species. Newtown Square, PA: US Department of Agriculture, Forest Service, Northeastern Research Station.

Keith H, Barrett D, Keenan R. 2000. Review of allometric relationships for estimating woody biomass for New South Wales, the Australian Capital Territory, Victoria, Tasmania, and South Australia. National Carbon Accounting System Technical Report no. 5b. Canberra: Australian Greenhouse Office.

Malimbwi RE, Temu AB. 1984. Volume functions for Pterocarpus angolensis and Julbernadia globiflora. *Journal of the Tanzania Association of Foresters* 1: 49–53.

Malimbwi R, Solberg B, Luoga E. 1994. Estimation of biomass and volume in miombo woodland at Kitulangalo Forest Reserve, Tanzania. *Journal of Tropical Forest Science* 7: 230–242.

Malimbwi RE, Shemwetta DTK, Zahabu E, Kingazi SP, Katani JZ, Silayo DA. 2005. Liwale district forest inventory report. Unpublished report. Dar es Salaam: Ministry of Natural Resources and Tourism.

Milledge S, Kaale B. 2003. Bridging the gap: linking timber trade with infrastructure development and poverty eradication efforts in southern Tanzania. Dar es Salaam: TRAFFIC East/Southern Africa.

Mugasha WA, Bollandsås OM, Eid T. 2013a. Relationships between diameter and height of trees in natural tropical forest in Tanzania. *Southern Forests* 75: 221–237.

Mugasha WA, Eid T, Bollandsås OM, Malimbwi RE, Chamshama SAO, Zahabu E, Katani JZ. 2013b. Allometric models for prediction of above-and belowground biomass of trees in the miombo woodlands of Tanzania. *Forest Ecology and Management* 310: 87–101.

Munishi PK, Temu PR-A, Soka G. 2011. Plant communities and tree species associations in a Miombo ecosystem in the Lake Rukwa basin, southern Tanzania: Implications for conservation. *Journal of Ecology and the Natural Environment* 3: 63–71.

Overman JPM, Witte HJL, Saldarriaga JG. 1994. Evaluation of regression models for above-ground biomass determination in Amazon rainforest. *Journal of Tropical Ecology* 10: 207–218.

Özçelik R, Brooks JR, Diamantopoulou MJ, Wiant HV Jr. 2010. Estimating breast height diameter and volume from stump diameter for three economically important species in Turkey. *Scandinavian Journal of Forest Research* 25: 32–45.

Parresol BR. 1993. Modeling multiplicative error variance: an example predicting tree diameter from stump dimensions in baldcypress. *Forest Science* 39: 670–679.

Philip MS. 1983. *Measuring trees and forests: a textbook written for students in Africa*. Dar es Salaam: University of Dar es Salaam.

Picard N, Saint-Andre L, Henry M. 2012. *Manual for building tree volume and biomass allometric equations: from field measurement to prediction*. Rome: Food and Agriculture Organization of the United Nations, Montpellier: Centre de Coopération Internationale en Recherche Agronomique pour le Développement.

R Development Core Team. 2013. R: A Language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org [accessed 12 November 2013].

Segura M, Kanninen M. 2005. Allometric models for tree volume and total aboveground biomass in a tropical humid forest in Costa Rica. *Biotropica* 37: 2–8.

Ter-Mikaelian MT, Korzukhin MD. 1997. Biomass equations for sixty-five North American tree species. *Forest Ecology and Management* 97: 1–24.

Thomas P, Packham J. 2007. *Woodland and forest ecology: Description, dynamics and diversity*. Cambridge: Cambridge University Press.

URT (United Republic of Tanzania). 1998. National Forest Policy. Dar es Salaam: Forestry and Beekeeping Division, Ministry of Natural Resources and Tourism.

URT. 2002. The Forest Act, no. 14 of 7th June 2002. Dar es Salaam: Ministry of Natural Resources and Tourism, the United Republic of Tanzania.

URT. 2010. National Forest Resources Monitoring and Assessment of Tanzania (NAFORMA) field manual – Biophysical survey. NAFORMA document M01–2010. Dar es Salaam: Forestry and Beekeeping Division, Ministry of Natural Resources and Tourism.

Vanclay JK. 1994. Modelling forest growth and yield: applications to mixed tropical forests. School of Environment, Science and Engineering Papers. Lismore: Southern Cross University.

Vieilledent G, Vaudry R, Andriamanohisoa SF, Rakotonarivo OS, Randrianasolo HZ, Razafindrabe HN, Rakotoarivony CB, Ebeling J, Rasamoelina M. 2012. A universal approach to estimate biomass and carbon stock in tropical forests using generic allometric models. *Ecological Applications* 22: 572–583.

Zahabu E. 2008. Sinks and sources: a strategy to involve forest communities in Tanzania in global climate policy. PhD thesis, University of Twente, The Netherlands.

Zhou X, Hemstrom MA. 2010. *Timber volume and aboveground live tree biomass estimations for landscape analyses in the Pacific Northwest*. General Technical Report PNW-GTR-819. Portland: US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Zianis D, Muukkonen P, Makipaa R, Mencuccini M. 2005. Biomass and stem volume equations for tree species in Europe. *Silva Fennica Monographs* 4. Helsinki: Finnish Society of Forest Science, Finnish Forest Research Institute.

Received 8 December 2013, revised 12 May 2014, accepted 26 May 2014

# PAPER II

# Modelling aboveground forest biomass using airborne laser scanner data in miombo woodlands of Tanzania.

Ernest William Mauya[1], Liviu Theodor Ene[1], Ole Martin Bollandsås[1], Terje Gobakken[1], Erik Næsset[1], Rogers Ernest Malimbwi[2] and Eliakimu Zahabu[2]

[1] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

[2] Department of Forest Mensuration and Management, Sokoine University of Agriculture, P.O.Box 3013, Morogoro, Tanzania.

**Abstract**

**Background**

Airborne laser scanning (ALS) has emerged as one of the most promising remote sensing technologies for estimating aboveground biomass (AGB) in forests, given its ability to provide canopy height information that is highly correlated with AGB. Use of ALS data in area-based forest inventories relies on the development of statistical models that relates AGB and metrics derived from ALS. Such models are firstly calibrated on a sample of corresponding field- and ALS observations, and then used to predict AGB over the entire area covered by ALS data. Several statistical methods, both parametric and non-parametric, have been applied in ALS-based forest inventories, but studies that compare different methods in tropical forests in particular are few in number and less frequent than studies reported in for example temperate and boreal forests. In this study, we compared linear mixed effects model (LMM) and *k*-nearest neighbor (*k*-NN).

**Results**

The results showed that the prediction accuracy obtained when using LMM was slightly better compared to the *k*-NN approach. Relative root mean square errors from the cross validation was 46.8% for the LMM and 58.1% for the *k*-NN. Post-stratification according to vegetation types improved the prediction accuracy of LMM more as compared to post-stratification by using land use types.

**Conclusion**

Although there were differences in prediction accuracy between the two methods, their accuracies indicated that both of the two methods have potentials to be used for estimation of AGB using ALS data in the miombo woodlands. Future studies on effects of field plot size and the errors due to allometric models on the prediction accuracy are recommended.

**Key words:** Parametric models, prediction accuracy, non-parametric models, LMM, k-NN, sampling design.

## Background

Estimation of aboveground biomass (AGB) in tropical forests is important for generating information needed for sustainable forest management and understanding the contribution of tropical forests in the global carbon cycle. Particularly in the latter context, estimates of AGB are needed as a primary variable for establishing the increments or decrements in carbon stored in tropical forests, which is typically converted from AGB by using a factor of 50% or less [1]. In the recent decade, Reducing Emissions from Deforestation and Degradation (REDD+), a program under United Nations Framework Conventions on Climate change, has motivated large-scale forest carbon inventories in tropical forests. REDD+ aims to provide positive incentives for developing countries for initiating activities related to reduction of greenhouse gas emissions, sustainable management of forests, and enhancement of forest carbon stock [2]. Unlike other conservation projects REDD+ is result based, which means that financial benefits relies on forest carbon stock changes that are measured, reported, and verified (MRV). Thus, establishment of effective MRV systems that complies with the guidelines of Intergovernmental Panel on Climate Change , is considered as an integral part of REDD+ implementation [3].

In Tanzania, the National Forestry Resources Monitoring and Assessment (NAFORMA), which is the national forest inventory of Tanzania, has established a total of 30,773 field plots distributed across mainland Tanzania [4]. NAFORMA is expected to be used to produce AGB data for the national forest carbon MRV system necessary for the implementation of REDD+ activities in Tanzania [5, 6]. However, being a field-based inventory, estimates of parameters related to AGB and AGB changes derived from NAFORMA data are not expected to be sufficiently precise to meet the accuracy requirements for a REDD+ MRV system. Therefore, the use of remotely sensed data as auxiliary information is considered as an option towards the development of a cost-efficient MRV system in the country.

Estimation of AGB using remotely sensed data in most of the tropical countries has been carried out using products from optical satellite sensors and synthetic aperture radar (SAR) satellite sensors. However, their accuracy are limited by a loss of sensitivity with increasing AGB, commonly known as "saturation", which occurs at relatively low AGB densities for optical data and somewhat higher for SAR data, for example around 100-250 Mg ha$^{-1}$ for L-band radar systems [7]. Airborne laser scanning (ALS) has recently received a great deal of scientific and operational attention because it can overcome the saturation problems of optical and SAR data. Its potential has previously been reported in particular in the Nordic countries where it has been used operationally for management inventories for almost 15 years [8]. Recently promising results from tropical forests [ e.g. 9, 10] have also been reported, which have increased the interest of using ALS for REDD+ MRV purposes. Apart from higher saturation level, ALS has the ability to directly measure vertical canopy profiles and also provides canopy height information which is highly correlated with forest AGB [11].

Large scale AGB assessments with ALS, however, remains to be a challenge due to logistics, cost, and data volume involved if wall-to-wall coverage is to be applied [12]. For such situations, a systematic sampling approach using ALS as a strip sampling tool is a viable option [13]. Within this approach, collection of ALS measurements is done along individual flight lines that cover only a small portion of the area of interest. The flight lines are aligned with a network of ground plots [13] which allow the development of statistical models relating the ground reference AGB to metrics derived from coincident ALS data. These models are then used to predict AGB over the entire area covered by ALS strips, and subsequently these predictions are used for final estimation of AGB for the area of interest using either design-based model-assisted or model-dependent inferential frameworks [ e.g. 14, 15]. Thus, the quality of the AGB estimates produced by ALS-based inventories relies heavily on the development and application of predictive AGB models.

A review study by Fassnacht et al. [16] has shown that the most common prediction methods in ALS-based forest inventories are ordinary least square regression, support vector machines, nearest neighbor-based methods (i.e. *k*-NN and *k*-MSN), and random forest. Of all the methods, ordinary least square regression with stepwise variable selection has been most frequently used for building models between field measurements and ALS metrics [17]. The main adavantage of using this type of methodology is the simplicity and clarity of the resulting models [18], especially when the relationship between AGB and the ALS metrics is almost linear. However, fitting and applicability of ordinary least square regression models rely on a number of basic assumptions in relation to the residual distribution: independence, normality and constant variance [19]. The assumptions are barely taken into account in most studies [20], especially when dealing with the data that are collected from complex field survey designs that involve clustered observations, repeated measurements, longitudinal measurements, and blocked data. Ignoring the model assumptions when fitting ordinary least square regression models, might lead to spatially correlated errors and biased standard error estimates for model parameters, and consequently, invalid significance tests [21].

Linear mixed effects models (LMMs) offer a modeling and prediction method that is very effective on clustered or spatially correlated data [22, 23]. In addition to accounting for covariates through fixed parameters as in ordinary least square regressions, mixed effects models can also account for various sources of heterogeneity and randomness in the data caused by known and unknown factors by means of random parameters. Application of LMMs are however limited in ALS-based inventories as compared to other prediction methods [24].

Non-parametric approaches, such as *k*-nearest neighbor (*k*-NN) has also been considered as an alternative to ordinary least regression since they do not rely on any distributional assumptions of the data [25, 26]. Thus, *k*-NN is a highly relevant alternative to deal with non-linear and possibly diverse relationships between independent and dependent variables.

Furthermore, like other nearest neighbor techniques, *k*-NN allows for both univariate and multivariate predictions of continuous and categorical variables [27, 28]. In forest inventory applications, *k*-NN approaches have been frequently applied in model-dependent frameworks with good results [29] and have also been used for mapping of various forest attributes [30, 31]. Several studies [ e.g. 32, 33, 34] have compared the performance of *k*-NN with ordinary least square regression (OLS) models in temperate and boreal forests, but few studies have compared LMMs with *k*-NN, especially in the context of the ALS-based inventory. Of particular interest is application and validation of such techniques in the tropical dry forests of Africa where application of statistical methods commonly used in ALS-based forest inventories still are limited compared with temperate and boreal forests. Given the growing potential of the use of national forest inventory data and ALS auxiliary information for supporting REDD+ activities in tropical forests [ e.g. 35, 6], it is important to explore modeling methods that fully utilize the attributes of design as a fundamental step towards reliable and accurate estimation of AGB using ALS.

Irrespective of the method used, stratification and post-stratification have been considered as effective tools for improving precision of estimates in ALS-based inventories [36]. Stratification according to forest age and/or site quality is commonly used in boreal forests [e.g. 37, 38]. In highly heterogeneous tropical forests, stratification/post-stratification based on vegetation types have been considered as a viable and practical option [ 39]. However, due to practical limitations, few studies have attempted to assess the effects of stratification and post-stratification on the prediction accuracy and thus on final estimates in tropical forests. For example in most of the previous studies, only a limited number of field plots have been available for AGB modeling due to issues such as accessibility and cost. Thus, stratification or post-stratification of a survey have not been regarded as viable since it could lead to even smaller sample sizes per stratum, making it difficult to fit reliable statistical models for each class [40]. In such situation, most of the previous studies opted to combine sample plot data across classes like for example vegetation types, and thus ignoring the effect of vegetation types and associated information.

The present study was conducted in the tropical forests of southern Tanzania which is mainly dominated by miombo woodlands and some areas with forest, cultivated land, and other vegetation types. Miombo woodlands occupy a substantial area of forest land in Tanzania (92%) [4] and it extends in other six countries in sub-Saharan Africa, including Angola, Zimbabwe, Zambia, Malawi, Mozambique and Democratic Republic of Congo [41]. From a global perspective, miombo woodlands have received a considerable attention in the recent decade given its potential to act as a reservoir of belowground and aboveground carbon stocks [42]. Biodiversity is also significant in the miombo woodlands with an estimate of 8500 species of higher plants and more than half of them are endemic [41]. Application of ALS in such areas represents the typical challenge that would be expected when using ALS data for modelling AGB in tropical forests with a high number of species, and diverse vegetation and land use types. The main objective of the study was to assess the performance of parametric and non-parametric methods for modeling and prediction of AGB using ALS data. As a secondary objective, we also assessed the effects of post-stratification by vegetation and land use types on the prediction accuracy of the parametric models.

## Results

### Performance of the parametric and non-parametric methods

The OLS model with square root transformed response variable was selected for building up LMMs. The model contained eight explanatory variables consisting of both height percentiles and the canopy density metrics selected using the best subset procedure. The OLS model showed cluster effects on the residual distributions as illustrated in Figure 1. Some clusters displayed residuals that were above and some below the zero line, indicating that cluster effects should be accounted for in the modelling. Comparison of the OLS model (Model1) and the LMM (Model 2) using likelihood ratio test (Table 1) suggested a statistically significance difference ($p < 0.001$) between the two models. Model 2 was considered to have better fit with low value of AIC as compared to Model 1.

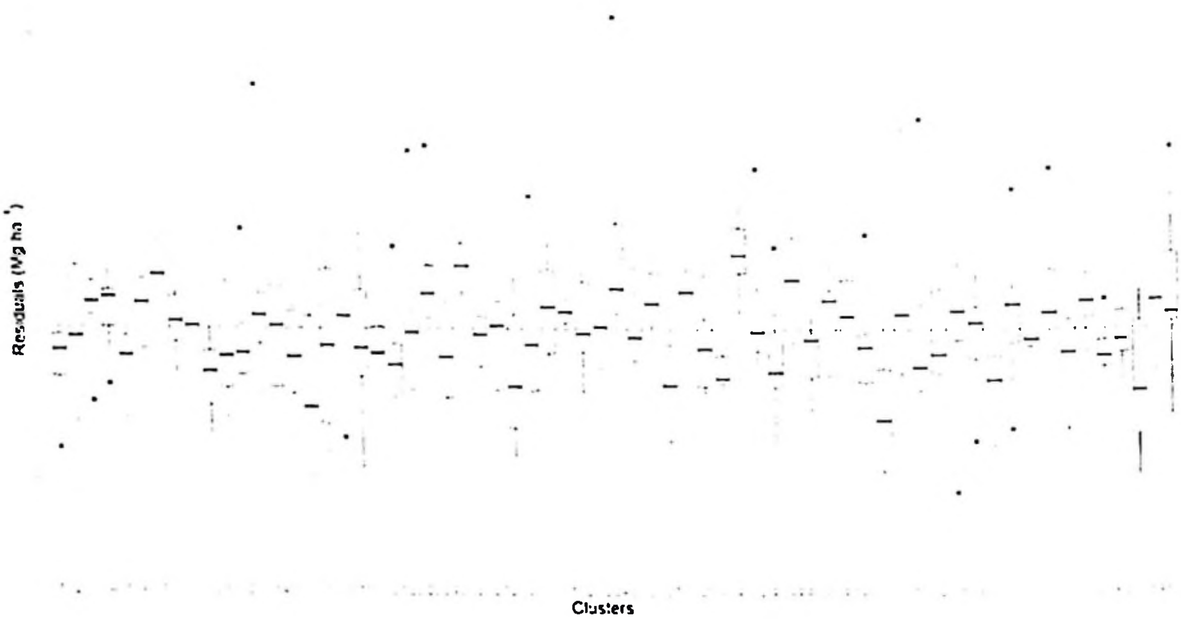



Figure 1. Box plot of residuals in the OLS model. The y-axis shows the value of the residuals and the horizontal axis the cluster.

Table 1: Results of likelihood ratio test comparing the initially selected models with models incorporating variance structure at the cluster level.

| Model [a] | AIC | Test | L.Ratio | *p*-value |
|---|---|---|---|---|
| Model1 | 1887 | | | |
| Model2 | 1873 | 1vs 2 | 11.02 | 0.001 |
| Model3 | 1849 | 2 vs 3 | 152.16 | <0.0001 |

[a] Model 1 = OLS model fitted with *gls function*. Model 2= LMM. Model 3= LMM with varIdent (form=cluster).

Re-fitting Model 2 with different correlation structures (i.e., spatial autocorrelation functions, and compound symmetry correlation structures), did not significantly improve model fit. The AIC values did not improve compared to the model without the autocorrelation functions (i.e., Model 2), and the likelihood ratio test indicated that there was no significance difference between the models with and without correlation structure ($p > 0.05$). This may also indicate that there is no spatial autocorrelation of the residuals within the clusters. Modelling the residual variance at the cluster level by using variance function (*varIdent*) improved the model performance as measured by the AIC. The likelihood ratio test indicated a statistically significant difference ($p < 0.0001$) between Model 2 and the Model 3 (Table 1). The standard errors of the parameters for Model 3 were smaller compared to the other models (Table 2). The quality of Model 3 was further analyzed by comparing an intercept model of Model 3 residuals and a similar model with a random intercept by means of the likelihood ratio test. The test indicated that the two models were not statistically significantly different from each other ($p > 0.05$), implying that Model 3 has successfully accounted for the dependency and heteroscedasticity in the data and thus the residuals can be considered as independent.

Table 2: Parameter estimates and standard errors of the tested models.

| Covariate [a] | Model 1 Estimate | Standard error | Model 2 Estimate | Standard error | Model 3 Estimate | Standard error |
|---|---|---|---|---|---|---|
| Intercept | 0.8630 | 0.2520 | 0.7385 | 0.2848 | 0.6702 | 0.2238 |
| PF70 | 0.3481 | 0.0338 | 0.3515 | 0.0345 | 0.3701 | 0.0273 |
| TF0 | 3.7779 | 0.5713 | 3.8436 | 0.6003 | 4.4363 | 0.4829 |
| TF5 | 4.0742 | 1.2839 | 3.7574 | 1.2964 | 1.7055 | 1.0384 |
| TF8 | -6.2178 | 1.9462 | -6.5277 | 1.9310 | -4.1964 | 1.4833 |
| PL20 | 0.3191 | 0.1241 | 0.2684 | 0.1232 | 0.1844 | 0.1100 |
| PL30 | -0.2295 | 0.1110 | -0.1845 | 0.1096 | -0.1059 | 0.0951 |
| TL7 | -12.4990 | 4.0241 | -11.9814 | 4.0086 | -6.8397 | 2.7174 |
| TL8 | 18.3840 | 5.5361 | 19.5302 | 5.4729 | 10.8928 | 3.7172 |

[a] PF70 = Percentile of the first echo canopy heights for 70% (m); PL20 and PL30 = Percentiles of the last echo canopy heights for 20% and 30% (m). TF0, TF5, TF8 = Canopy densities corresponding to the proportion of first echoes above fraction #0 (1.3m), #5 and #8 (see text); TL7, TL8 = Canopy densities corresponding to the proportion of last echoes above fraction #7 and #8 (see text).

Table 3: Predictors, pseudo R-squared ($R^2$), and relative root mean square error from the cross validation (RMSEcv%) of the two prediction methods.

| Prediction method | Predictors [a] | $R^2$ | RMSEcv% |
|---|---|---|---|
| LMM (Model 3) | PF70, TF0, TF5, TF8, PL20, PL30, TL7, TL8 | 0.69 | 46.8 |
| *k*-NN (k=5) | PL80, TF0, TL2, PF80, TL4, TL8, PL40, TL5, TF2, TL7, TL6, PF70, MaxF, PL90, PL50, MeanL, TF7, PF10, PL60, TF3 | 0.63 | 58.1 |
| *k*-NN (k=10) | PL80, TF0, TL2, PF80, TL4, TL8, PS40, TL5, TF2, TL7, TL6, PF70, MaxF, PL90, PL50, MeanL, TF7, PF10, PL60, TF3 | 0.58 | 55.9 |

[a] PF10, PF70, PF80 = Percentiles of the first echo canopy heights for 10%, 70%, and 80 (m), PL20, PL30, PL40, PL50, PL60, PL90 = Percentiles of the last echo canopy heights for 20%, 30% , 40%, 50%, 60% and 90% (m); TF0, TF2, TF3, TF5, TF7, TF8 = Canopy densities corresponding to the proportion of first echoes above fraction #0 (1.3m), #2, #3, #5, #7 and #8;TL2, TL4, TL6, TL7, TL8, = Canopy densities corresponding to the proportion of last echoes above fraction #2, #4 , #6, #7, and #8; MaxF and MeanL = Maximum and Mean of the canopy height distributions of the first and last echoes respectively

The *k*-NN method fitted with two values of *k* (i.e., $k = 5$ and $k = 10$) were tested. The *k*-NN imputation with $k = 10$ was having relatively lower RMSEcv% values compared to imputation with $k = 5$. We tested the dependency and heteroscedasticity of the residuals obtained from *k*-NN. Two models were compared using the likelihood ratio test, i.e., a residual intercept model and a random intercept model. The results from the likelihood ratio test showed that there was no statistically significant differences between the two models. Comparing the results of the best parametric model (i.e., Model 3) and the non-parametric (i.e., $k$ = 10) (Table 3). our results suggest that the parametric models performed well in our dataset as indicated by both $R^2$ and RMSEcv%. Graphical illustrations for the performances of the two methods are presented in Figure 2.

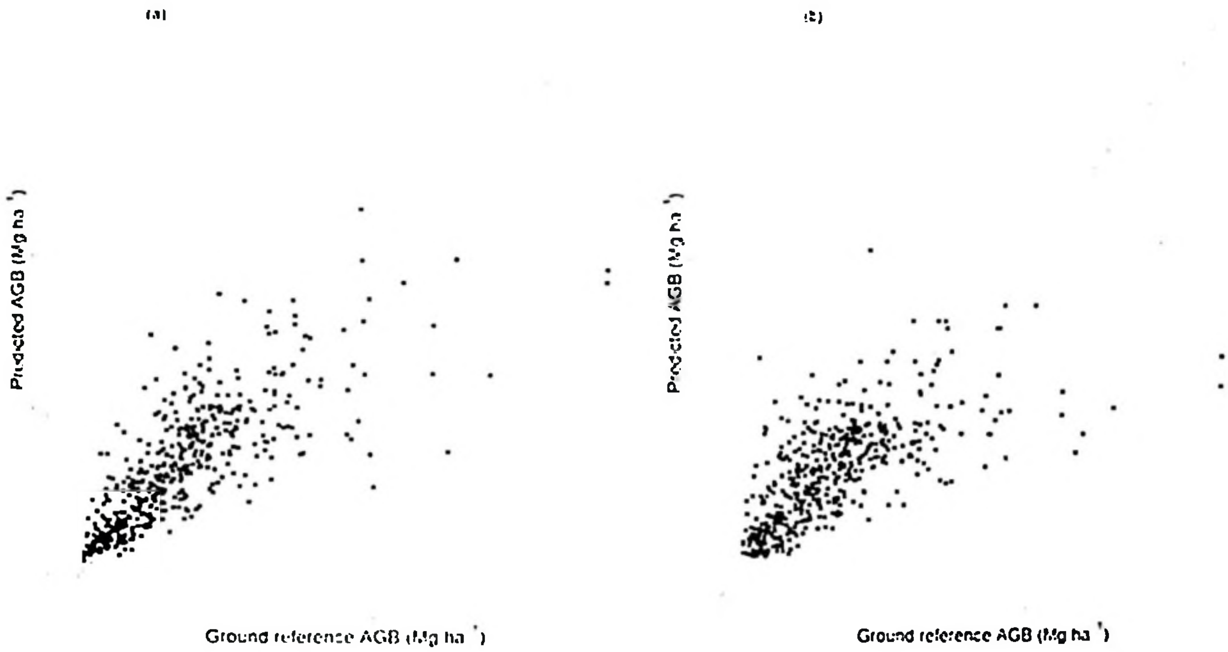



Figure 2. Relationship between ground reference and predicted AGB for LMM (a) $k$-NN ($k$ = 10) (b).

### *Effect of post-stratification on prediction accuracy of the parametric models*

To account for the effect of post-stratification on model accuracy, we assessed the performance of parametric model (i.e., Model 3) on different vegetation and land uses types (Table 4). The results indicated that there were variations in the prediction accuracy of the model across the categories. The RMSE% and RMSEcv% of Model 3 (i.e., non-post-stratified model) varied from one category to another, lower values of RMSE% and RMSEcv% were reported for vegetation types as compared to land use types.

Separate random intercept models were fitted for each of the categories (Table 5) and compared with the non-post-stratified model presented in the previous section. Generally, the

RMSEcv% for the post-strata models were relatively smaller compared to the values obtained when evaluating the non-post-stratified model across respective post-strata. The accuracy of the post-strata models varied in terms of model fits (i.e., $R^2$) and RMSEcv% depending on the vegetation and land use types (Table 5). Graphical plots in Figures 3 and 4 illustrate the performance of the post-strata models.

Table 4: Performance of the parametric model (Model 3) for different vegetation and land use types.

| Category[a] | RMSE (Mg ha$^{-1}$) | RMSEcv (Mg ha$^{-1}$) | RMSE% | RMSEcv% |
|---|---|---|---|---|
| Vegetation type | | | | |
| Forest | 22.7 | 25.1 | 25.7 | 28.4 |
| Woodlands | 29.4 | 31.6 | 44.3 | 47.7 |
| Other cover types | 37.6 | 38.5 | 80.2 | 82.1 |
| Land uses type | | | | |
| Production and protection forests | 30.4 | 31.1 | 40.2 | 41.1 |
| Wildlife reserves | 25.5 | 26.3 | 50.2 | 52.0 |
| Agriculture and other land use types | 33.2 | 34.0 | 73.0 | 74.7 |

[a] Forest = land spanning more than 0.5 ha with trees that have heights of more than 5 m and a canopy cover of more than 10%. It does not include land that is predominantly under agricultural or urban land use. Woodland = forestland with less dense canopy cover compared to forest. Other cover types = all cover types that were neither forest nor woodlands. Production and protection forests = forest areas designated for protection of water (i.e catchment forests) and that designated for production of wood respectively. Wildlife reserves = forest areas designated for game reserves and game controlled areas. Agriculture and other land use types = areas designated primarily for a function other than production, protection or game reserves. Details descriptions of these categories are given in MNRT [62]

Table 5. Predictors, number of observations (n), pseudo R-squared ($R^2$), and relative root mean square error from the LOOCV (RMSEcv%) for separate LMM fitted for different vegetation and land use types.

| Category[a] | Predictors | n | $R^2$ | $RMSE_{cv}$% |
|---|---|---|---|---|
| **Vegetation type** | | | | |
| Forest | PF20, MaxL, MeanL, PL10, PL40, PL70, PL90, TL9 | 40 | 0.85 | 23.9 |
| Woodlands | PF70, TF0, PL20 | 391 | 0.63 | 45.3 |
| Other cover types | MeanF, PF20, PF60, TF5, MaxL, PL70 | 58 | 0.83 | 68.3 |
| **Land use type** | | | | |
| Production and protection forests | PF40, PF60, PF70, TF0, TF9, PL20 | 314 | 0.64 | 40.8 |
| Wildlife reserves | CVF, PF20, PF90, TF5 | 91 | 0.73 | 49.8 |
| Agriculture and other land uses | TF1, TF4, MaxL, MeanL, CVL, TL0, TL4 | 84 | 0.69 | 68.0 |

[a] **Forest** = land spanning more than 0.5 ha with trees that have heights of more than 5 m and a canopy cover of more than 10%. It does not include land that is predominantly under agricultural or urban land use. **Woodland** = forestland with less dense canopy cover compared to forest. **Other cover types** = all cover types that were neither forest nor woodlands. **Production and protection forests** = forest areas designated for protection of water (i.e. catchment forests) and that designated for production of wood respectively. **Wildlife reserves** = forest areas designated for game reserves and game controlled areas. **Agriculture and other land use types** = areas designated primarily for a function other than production, protection or game reserves. Details descriptions of these categories are given in MNRT [62]

[b] PF20, PF40, PF70, PF90 = Percentiles of the first echo canopy heights for 20%, 40%, 70% and 90 (m); PL10, PL20, PL70, PL90 = Percentiles of the last echo canopy heights for 10%, 20%, 70%, and 90% (m); TF0, TF1, TF4, TF5 = Canopy densities corresponding to the proportion of first echoes above fraction #0 (1.3m), #1, #4, and #5; TL0, TL4, TL6, TL7, TL8, = Canopy densities corresponding to the proportion of last echoes above fraction #0 (1.3m), #4, #6, #7 and #8, MeanF and MeanL = arithmetic mean of first or last echo laser canopy heights, respectively (m), MaxF and MaxL = maximum of first or last echo laser canopy heights, respectively (m); CVF and CVL = coefficient of variations for the first and last echo laser canopy heights respectively.

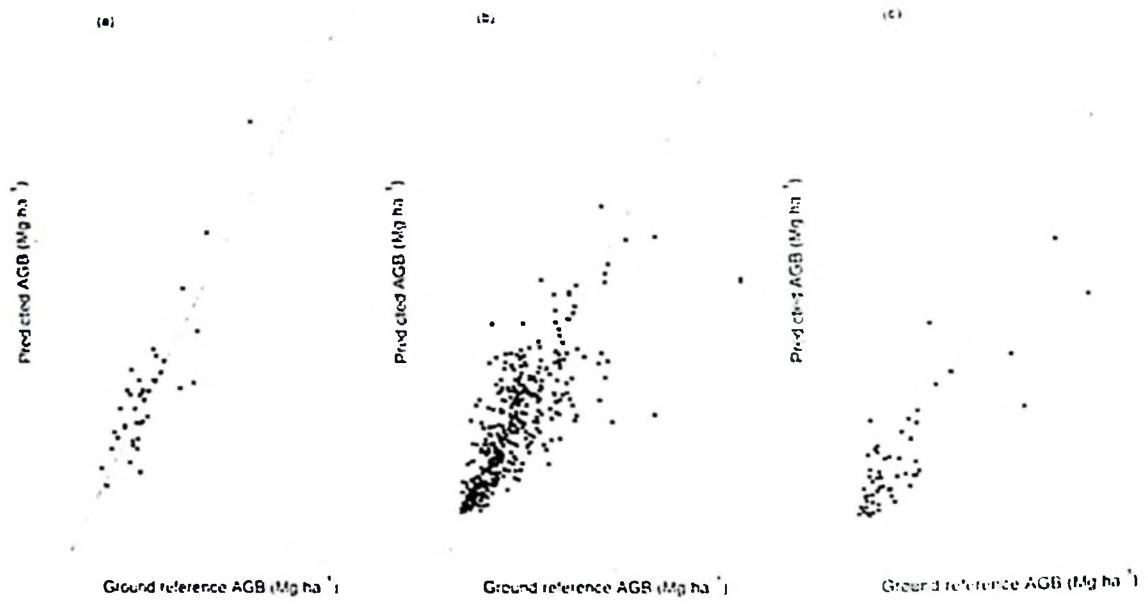



Figure 3. Relationship between ground reference and predicted AGB for LMM in forest (a) woodlands (b) other cover types (c).

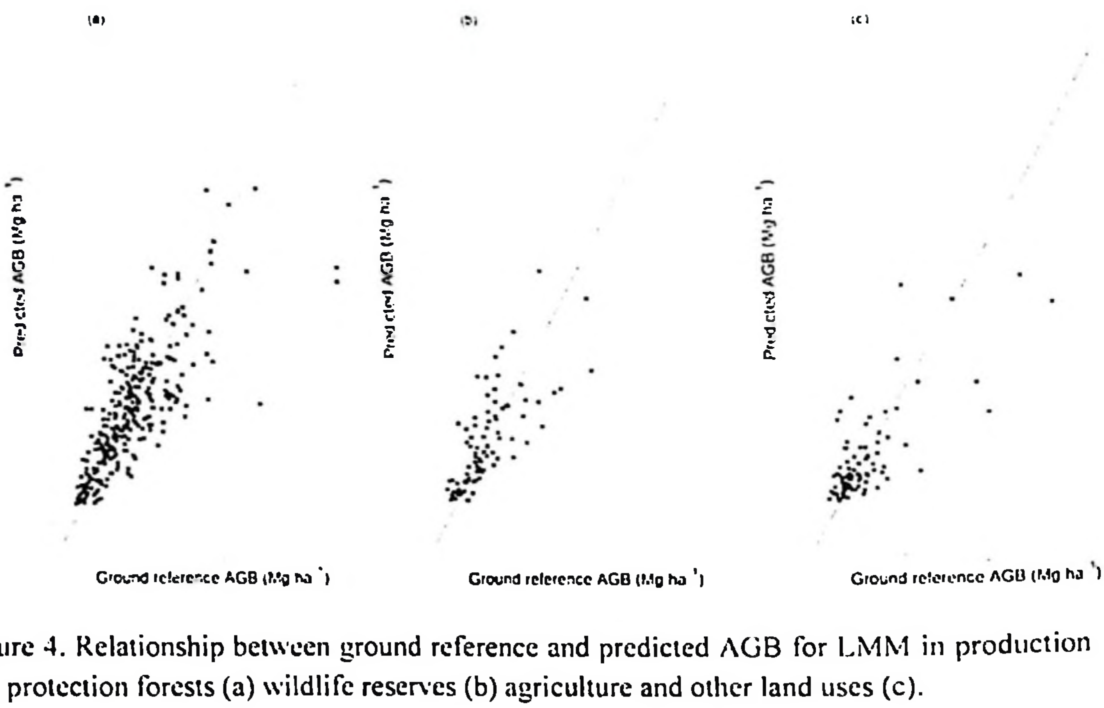



Figure 4. Relationship between ground reference and predicted AGB for LMM in production and protection forests (a) wildlife reserves (b) agriculture and other land uses (c).

### *Discussion*

This study aimed at comparing the performance of the parametric (LMMs) and non-parametric (*k*-NN) methods for predicting AGB using ALS data in the miombo woodlands of Tanzania. Effects of post-stratification by vegetation and land use types on the prediction accuracy of the parametric method were considered as the secondary objective. The findings from this study have demonstrated that both LMMs and *k*-NN are suitable methods for predicting AGB using ALS data. To our understanding this is one among the early studies that attempted to use ALS in the miombo woodlands of Tanzania. Thus the findings of this study open up methodological insights on the use of ALS as tool for AGB assessment in similar type of vegetation in sub-Saharan Africa.

Specifically the findings have shown that LMM is the best prediction method, both by allowing the specific field sampling design to be accounted for in the modeling, but also by having slightly higher prediction accuracy compared to *k*-NN. This is not surprising, and has been reported in most of the studies that have attempted to compare parametric and non-parametric methods in prediction of various forest attributes [e.g. 32, 43, 44]. However, the strength of *k*-NN is that it was able to account for the dependence and heteroscedasticity in the data. This indicates that it can reliably be used for estimation and making inference when deemed necessary, especially in the design-based framework of forest inventory [ e.g. 45, 46] with non-parametric based estimators (e.g. difference estimators).

Results based on the LMM illustrated that incorporating the cluster structure by using variance function in the model selection process can result in a model with better fit as supported by the likelihood ratio test. This implies that it is the between-cluster variability that should be considered when calibrating the ALS models using NAFORMA data, rather than the within-cluster variability. By modelling the residual variance per cluster through an appropriate variance structure (i.e., the *varIdent* structure) we were able to account for this variability in the data, which resulted in a model with low standard errors for the parameter estimates as compared to the other tested models (Table 2). Low standard errors of the parameter estimates (Table 2) indicates that the model is more efficient when predicting outside the sample. Furthermore, low standard errors of the parameter estimate is an important property in improving precision of the estimates, especially when making inference using model-based estimators [47] which theoretically relies on the quality of the model parameters.

Further evaluation of the best non-post-stratified model across post-strata, showed that there were variations in prediction accuracy across different vegetation and land use types. This could likely be attributed to the difference in stem diameter and the number of trees per unit area, which entirely affect the distribution of AGB and the characteristics of the ALS data in each of the post-strata. When fitting separate models by post-strata, our findings showed that that there was a slight gain in prediction accuracy compared to the use of non-post-stratified model in the respective post-strata. This might be due to the homogeneity of the respective category, which in turns improves the relationship between ALS metrics and the ground reference AGB. For example, in the post-strata such as forest or production and protection forests, where the distribution of AGB is characterized by trees with high canopy cover and more uniform stems, the $RMSE_{cv}$% were relatively lower compared other categories. On the hand, the higher $RMSE_{cv}$% values in the agricultural and other land uses might be attributed by the fact that most of the trees in this category are scattered with sparse canopy, and the tree crowns are smaller and some appear in a degraded form.

Although post-strata models performed well compared to non-post-stratified model, their practical application in the miombo woodlands poses a number of challenges when used for estimation and inference. Based on the sampling design described in this study, the use of post-strata models would require having thematic maps for the land use classes and vegetation types. Such maps are not trivial to produce, and our results (not presented) indicated that the classification accuracies vary substantially among these categories. Thus, since the difference between the non-post-stratified model and the post-strata models were modest, we would rather recommend the un-post-stratified model (which disregards the land use and vegetation types) to be more adequate for most applications that will involve large-scale AGB estimation supported by ALS data, at least until high quality thematic maps are made available.

Generally, the finding of our study in terms of model quality criterions such as $R^2$ and RMSEcv% for non-post-stratified and post-strata models, are in line with most of the published studies from tropical forests [48-51]. A majority of these studies reported $R^2$ ranging from 64% [52] to 90% [53]. Similarly, a study by Asner et al. [49] across four tropical regions in Panama, Peru, Madagascar, and Hawaii reported $R^2$ ranging from 0.68 to 0.85. Recently, a study from the tropical rainforest of Tanzania by Hansen, *et al.* [9] reported RMSEcv% ranging from 32.3% to 36.8% for models with different forms and different sets of predictor variables. However, direct comparison with these results should be taken with caution, due to wide range of variations existing in the tropical forests and different sample sizes and plot sizes used in different studies.

Even though we are convinced that our findings reflect the potential performance of ALS-based AGB models in dry tropical forest conditions, but there might be ways to further improve the quality of the models. For instance, the plot size used in this study was relatively smaller compared to what have been used by the studies reporting higher prediction accuracy [e.g. 54, 49]. Most of these studies have used plot sizes that are even twice as large as used in the current study. For example, Mauya et al. [55] reported a decrease in RMSEcv% from 63.6% to 29.2% for plot sizes ranging from 200 $m^2$ to 3000 $m^2$ in a high-biomass rainforest. The increase in prediction accuracy for studies based on larger plots might be attributed to the so-called spatial averaging of the errors, because both the field observations and the ALS data capture more of the spatial variation and are closer to the average value [e.g. 56, 9]. Furthermore, the relative and negative influence of plot positioning error on the prediction accuracy is reduced for the larger plot sizes because the overlap between the field- and ALS-data becomes larger as plot size increases [57]. In addition, plot boundary effect which have potential to cause discrepancies between field and ALS-based measurements, is reported to be relatively lower for the larger plots as compared to the smaller plots [55].

The concentric design of the field plots used in the current study also introduced errors in the relationships between AGB estimated on the plots and the ALS data. With this design, small trees are measured only in the center of a field plot while the largest trees are measured across the entire plot. However, smaller trees are also found in the outer part of a plot, and these trees will be measured by the laser but not recorded in the field data. Measuring all trees across a plot would clearly improve model fit. However, this study focused on the already existing design and data established by NAFORMA, thus it was also important to demonstrate how the NAFORMA data would be used with ALS auxiliary information. Similarly, miombo woodlands are dominated by a lower herbaceous layer of the vegetation which was not accounted for in the field measurements but were captured by the ALS data. Although a threshold of 1.3 m was applied to the ALS data to define the "canopy" layer, it is likely that the ALS data contain height observations reflected from vegetation not recorded by the tree measurements. This has certainly introduced additional errors in the models and reduced their performance. Lastly, it should be mentioned that the errors associated with the allometric models used to compute AGB (which were ignored in this study) will tend to affect the accuracy of the ALS-based prediction models. A general model by Mugasha et al. [58] that combines all the tree species was used for computation of AGB on the ground plots. Given the high number of tree species it is likely that the uncertainty of the field-based reference values is substantial. To what extent this error has effect on the prediction accuracy of the ALS-based models, is not yet known in the miombo woodlands and should be focused in future research.

### *Conclusions*

To conclude, the study has demonstrated that prediction of AGB using ALS data can reliably be done in the miombo woodlands of Tanzania. The comparison of the prediction methods have shown that LMM is the most appropriate method for AGB prediction using ALS data, as measured by the results from the RMSEcv% but also by considering its strength of accounting for the complex sampling design of the NAFORMA program. The prediction accuracy of *k*-NN was relatively smaller compared to LMM, yet it can be used when there is a need for using non-parametric method. Post-stratification by vegetation types seemed to favor the prediction accuracy compared to land use types. However, the non-post-stratified model has relatively more advantages due to its versatility and practical limitations of using post-strata models. Thus, we suggest using LMM (i.e., Model 3) that combines all the post-strata for applications involving large-scale AGB estimation in the future. Lastly, our study identified important knowledge gaps and directions of future research, such as assessment of the effects of field plot size and use of on-plot protocols which is based on complete census of all the trees in a plot rather than a sample according to tree size. Finally, a better understanding and quantification of the effects of allometric model errors on overall uncertainties of ALS-based models and AGB estimates is a fundamental topic for future studies.

## Methods

### Study area

The study area is located in Liwale district (9°54'S, 37°38'E) (Figure 5a), Lindi region, Tanzania, and has a total size of 15,867 $km^2$ (Figure 5b). The mean annual temperature of Liwale district ranges between $20^{0}$C and $30^{0}$C. Rainfall pattern is bi-modal with a dry season from June to October. A short rain period starts from late November and lasts in January. There is dry spell in February followed by a longer wet season which lasts from March to May. The mean annual rainfall ranges from 600 mm to 1000 mm [59]. The study area contain typical miombo flora of high trees with shrubs and grasses on the forest floor. In general, the area is characterized by high species diversity associated with typical miombo tree species such as *Brachystegia sp., Julbernadia sp.*, and *Pterocarpus angolensis*.

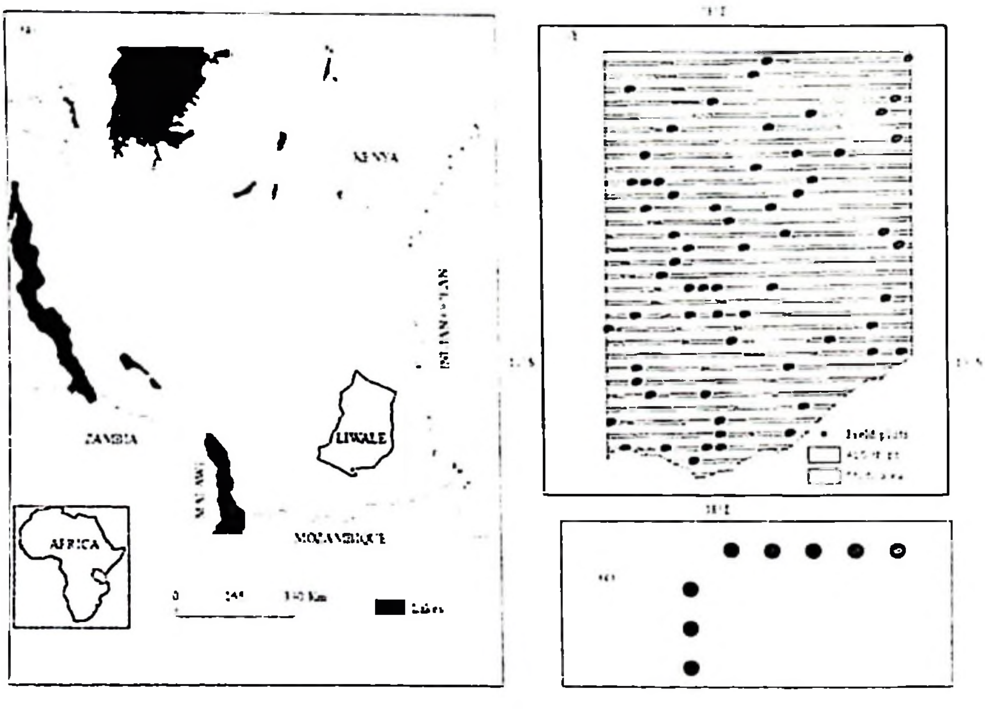


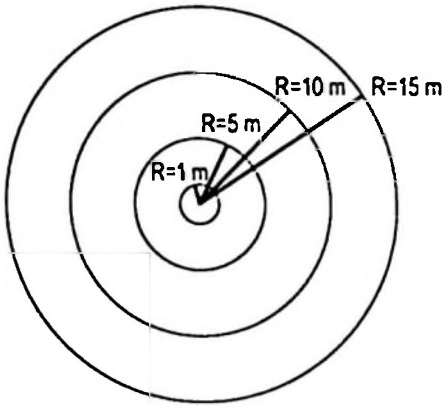

Figure 5. Map showing location of the study area in Liwale, southern Tanzania (a), distribution of clusters of field plots and ALS strips in the study area (b) and the cluster structure with distribution of individual field plots within clusters (c).

## Sampling design

The field plots used in this study were initially established by NAFORMA in 2011. The sampling design used by the NAFORMA is double-sampling for stratification which was designed based on a simulation study described by Tomppo et al. [6]. The first-phase sample consists of clusters of plots on a 5×5 km grid over mainland Tanzania. The first-phase clusters were stratified based on predicted growing stock, time consumption for cluster measurements and slope of the terrain [ 6]. All together, the first-phase clusters that contain 6 to 10 plots (Figure 5c) per cluster were assigned to 18 pre-defined strata. The second-phase samples were systematically selected from the first phase sample, with different sampling intensities in each of the 18 strata following an optimal allocation procedure [60] with cost functions tailored for each stratum. Greater sampling intensity was allocated to strata with large predicted growing stock and smaller sampling intensity to strata with small predicted growing stock. Only the clusters selected during the second phase of sampling were measured in the field. The distance between field plots within a cluster was 250 m, while the distance between clusters varies from the shortest possible distance (5 km) to 45 km.

## Field measurements

NAFORMA field plots were revisited during the first quarter of 2012. The aim of the field work was to accurately record the positions of the field plots and to update the field information to have temporal consistence between field measurements and the time for acquisition of ALS data. Measurements on the plots followed the same procedure used by NAFORMA in 2011. The circular plots of 15 m radius were identified. Diameter at breast height (*dbh*) was measured using caliper or diameter tape following the lower *dbh* thresholds in accordance with the concentric plot design described in Figure 6.




1) Within 1 m radius: all trees with *dbh* > 1 cm were recorded.

2) Within 5 m radius: all trees with *dbh* > 5 cm were recorded.

3) Within 10 m radius: all trees with *dbh* > 10 were recorded.

4) Within 15 m radius: all trees with *dbh* > 20 cm were recorded.

Fig. 6. NAFORMA concentric sample plot design.

Species names were recorded for every tree measured for *dbh*. Every fifth tree in the cluster was selected as sample tree and measured for height using Suunto hypsometers. The heights of the remaining trees were predicted using diameter-height models developed based on the sample trees. Differential Global Navigation Satellite Systems were used to calculate the coordinates of the center point of each sample plot. Two Topcon Legacy 40-channels dual frequency receivers observing both pseudo-range and carrier phase of the Global Positioning System (GPS) and the Global Navigation Satellite System (GLONASS) were used as rover (on the plot) and base station, respectively. Based on the positional standard errors reported by Pinnacle [61], the estimated accuracy of the planimetric plot coordinates ranged from 0.004 to 1.334 m, with an average of 0.194 m.

Tree AGB was estimated using allometric models for miombo woodlands developed by [58]. The AGB estimates of the individual trees were then summed for each plot, and scaled to per-hectare values. The plots were grouped according to their respective stratum, land use, and vegetation types following the procedure described by NAFORMA in MNRT [62]. In this study, we narrowed the land use classes and the vegetation types described in MNRT [62] into three categories to simplify the interpretation of the results, but also to have enough samples for each category (Table 6). The land use classes were grouped into: 1) production and protection forests, 2) wildlife reserves, 3) agricultural and other land use types. Similarly, the vegetation types were grouped into: 1) forest, 2) woodlands, 3) other cover types.

Table 6. Summary of field data. Number of plots for the different strata and post-strata together with minimum, maximum, and mean ground reference AGB values with their corresponding standard deviation.

| Category [a] | Number of plots | AGB (Mg ha$^{-1}$) | | | |
|---|---|---|---|---|---|
| | | Minimum | Maximum | Mean | Standard deviation |
| Stratification | | | | | |
| Stratum 5 | 116 | 3.9 | 158.2 | 45.9 | 29.8 |
| Stratum 6 | 81 | 2.2 | 179.2 | 51.3 | 41.7 |
| Stratum 7 | 90 | 2.4 | 270.3 | 84.7 | 61.3 |
| Stratum 8 | 119 | 0.3 | 349.9 | 79.4 | 53.7 |
| Stratum 9 | 23 | 0.3 | 125.6 | 38.7 | 35.4 |
| Stratum 10 | 6 | 0.3 | 56.1 | 16.0 | 20.5 |
| Stratum 11 | 16 | 55.1 | 186.9 | 87.9 | 32.4 |
| Stratum 12 | 38 | 34.3 | 182.7 | 85.2 | 35.6 |
| Post-stratification | | | | | |
| Vegetation type | | | | | |
| Forest | 40 | 27.8 | 232.2 | 88.3 | 39.6 |
| Woodlands | 391 | 0.3 | 349.9 | 66.3 | 48.4 |
| Other cover types | 58 | 0.3 | 270.3 | 46.9 | 53.7 |
| Land use type | | | | | |
| Production and protection forests | 314 | 0.3 | 349.9 | 75.6 | 49.5 |
| Wildlife reserves | 91 | 0.3 | 190.6 | 50.8 | 39.0 |
| Agriculture and other land uses | 84 | | 270.3 | 45.5 | 48.1 |
| All | 489 | 0.3 | 349.9 | 65.8 | 49.2 |

[a] Stratum 1 to 6 refers to the strata to which field plots belongs as described in the text and elaborated by [6] .**Forest** = land spanning more than 0.5 ha with trees that have heights of more than 5 m and a canopy cover of more than 10%. It does not include land that is predominantly under agricultural or urban land use. **Woodland**= forestland with less dense canopy cover compared to forest. **Other cover types** = all cover types that were neither forest nor woodlands. **Production and protection forests**= forest areas designated for protection of water (i.e catchment forests) and that designated for production of wood respectively. **Wildlife reserves** = forest areas designated for game reserves and game controlled areas. **Agriculture and other land use types** = areas designated primarily for a function other than production, protection or game reserves. Details descriptions of these categories are given in MNRT [62].

**ALS data**

Acquisition of the ALS data was carried out along 32 parallel strips with an average width of 1374 m, which were systematically distributed over the study area in east-west direction. The ALS strips were spaced 5 km apart, following the NAFORMA 5×5 km grid. A Leica ALS 70 airborne laser sensor (Leica Geosystems AG, Switzerland), carried by a Cessna 404 aircraft, was used to acquire the data from 10 February to 7 March 2012. The measurements were acquired from an average flying altitude of approximately 1200 m above ground, at an average ground speed of 77.2 $ms^{-1}$. The scanning rate was 36.5 Hz and the instrument operated at a pulse repetition frequency of 193 kHz. The average point density was around 1.8 points $m^{-2}$.

Processing of the ALS data started with the classification of the ALS echoes into ground or vegetation echoes using the progressive irregular triangular network densification method [63, 64] implemented in the TerraScan software [64]. A triangular irregular network (TIN) was created using the ALS echoes classified as ground echoes. The heights above the ground surface were then calculated for all vegetation echoes by subtracting the TIN height at their respective xy-positions. Up to five echoes were registered per pulse, but we used only the three echo categories "single", "first of many", and "last of many". The "single" and "first of many" echoes were pooled into one dataset denoted as "first" echoes, and correspondingly, the "single" and "last of many" echoes were pooled into a dataset denoted as "last" echoes.

For each echo category, height distributions were first created as described by [65]. A height threshold of 1.3 m was applied in order to separate trees from falsely classified ground features and low vegetation. Then, heights at ten percentiles (0th, 10th, ..., 90th) of these height distributions were computed to represent canopy height distribution and labeled PF0, PF10, ..., PF90 (first echoes) and PL0, PL10, ..., PL90 (last echoes), respectively. Furthermore, measures of the canopy density were also computed for first and last echoes. The range between the lowest ALS canopy height (>1.3 m) and the 90th percentile height was divided into 10 layers of equal height. Canopy densities were then computed as the proportion of echoes above each layer to total number of first echoes and labeled TF0 (>1.3 m), TF1, ..., TF9. Density variables for the last echo distributions were calculated in the same way and labeled TL0, TL1, ..., TL9. Furthermore, mean (MeanF and MeanL), maximum (MaxF and MaxL) and coefficient of variation (CVF and CVL) of the canopy height distributions were also computed for both first and last echoes.

## Statistical analyses

### An overview

Three statistical techniques were used to develop relationships between the ground reference AGB and the ALS metrics. These included OLS, LMMs, and *k*-NN technique. The analyses were performed based on the five outlined steps below:

1. Candidate explanatory variables from the ALS metrics were selected and three OLS model forms relating ground reference AGB and ALS metrics were fitted and tested.
2. The best selected model form from step 1 was used to build LMM (i.e., random intercept model) with random effect at the cluster level (i.e., we added random effect to the OLS model). The OLS and the LMM were compared.
3. To account for spatial dependence within the clusters we introduced LMMs with different correlation structures and compared with the LMM (i.e., random intercept model) fitted in step 2.
4. LMM with variance structure at the cluster level was also fitted. The model was compared with the LMM fitted in step 2 using likelihood ratio testing. The best selected model (i.e., from step 1 to 4) was further evaluated using a cross validation procedure.

5. Finally, the $k$-NN imputations were fitted and compared with the best model selected from the procedure described above using measures of reliability based on cross validation.

**Parametric methods**

**Model development (OLS)**

OLS are among the most common methods for modeling and predicting AGB in ALS-based forest inventory. As part of the model development procedure, we first applied an automated approach to select candidate predictor variables using the "*regsubset*" function from the leaps package [66] in the R statistical software [67]. The "*regsubset*" regression perform an "all subsets" where all possible variable combinations are considered and ranked based on different scoring criteria (adjusted $R^2$, Mallow's $C_p$ statistics, BIC etc.) [68]. We used Mallow's $C_p$ statistics [69], which is a measure of the degree of over-fitting, calculated from the error sum of squares for the model under evaluation and the mean square error of the model including all available predictors. Mallow's $C_p$ is given by the formula

$$C_p = \frac{SSE}{MSE_{full}} - N + 2p \tag{1}$$

where $SSE$ is the error sum of squares of the reduced model with $p$ parameters (including the intercept), $MSE_{full}$ is the mean square error of the full model, and $N$ is the sample size. A combination of predictors that minimizes the Mallow $C_p$ over all possible subsets was considered as the best subset [19]. The variable selection was repeated for log-transformed variables and square root transformed response variable. Thus, three types of OLS models were finally fitted and tested. Of all the three model forms, square root transformation (equation 2) was selected as the best based on our initial test results (not presented), i.e.,

$$\sqrt{y_i} = \beta_0 + \beta_1 x_{i_1} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1 \ldots\ldots n \quad \varepsilon_i \sim N\,(0, \sigma_\varepsilon^2) \tag{2}$$

Where $y_i$ is the ground reference AGB of the $i$th sample plot, $x_{i1} \ldots\ldots\ldots, x_{ik}$ are the $k$ predictor variables (i.e ALS metrics). $\beta_{0,} \ldots\ldots \beta_k$ are the parameter estimates. $n$ is the number of sample plots and $\varepsilon_i$ is the plot level residuals.

**Model development (LMM)**

The sampling design employed by NAFORMA imposes a hierarchical data structure by which the field plots are nested within the clusters. LMM is in such a case considered as an ideal tool for development of predictive models [23, 70] that will account for dependence of the plots within the clusters. LMM consists of two main parts, fixed and random effects. The fixed effects are common to all subjects, while random effect parameters are specific to each subject [71]. In this study the predictor variables used in the OLS with square root transformed response variable from the previous section were used as the fixed effects and the cluster id was added to the model to account for the random effect part. The standard form of LMM as applied in this study is

$$\sqrt{y_{ij}} = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_k x_{ijk} + u_i + \epsilon_{ij} \tag{3}$$

$$i = 1, \ldots\ldots, n_i \quad j = 1, \ldots\ldots, m \quad u_i \sim N\big(0, \sigma_\mu^2\big) \qquad \epsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where $y_{ij}$ is the ground reference AGB of $i$th sample plot in the $j$th cluster.$x_{ij1}, \ldots\ldots\ldots\ldots, x_{ijk}$ are $k$ fixed effects. $\beta_0, \ldots\ldots\ldots\ldots, \beta_k$ are the fixed effects parameters, $n_i$ is the number of sample plots within the cluster j . $m$ is the number of clusters. We assumed that cluster level random effects $\mu_i$ were independent of the plot level residuals $\varepsilon_{ij}$ .

To evaluate the significance of the random effect, we re-fitted the OLS model using generalized least square *function* in order to compare the OLS with the LMM using the likelihood ratio tests, as described by [70].

To account for the non-constant variance and spatial autocorrelation that might not have been accounted for by the random effect, we further refitted the LMM with variance and correlations structures and compared with the LMM (i.e., the random intercept model). The details for this procedure is described below and elaborated further by [70].

**LMMs with correlation structures**

We fitted five different LMMs using maximum likelihood estimation (ML), each assuming different spatial autocorrelation structures, i.e., linear, ratio, exponential, spherical, and Gaussian. This was aimed at testing the effect of spatial autocorrelation to account for field plot proximity within the clusters. In addition to that, we also tested compound symmetry correlation structure, assuming that correlation among plots within a cluster is constant but might vary from one cluster to another. The LMMs that incorporate spatial autocorrelation and compound symmetry correlation structures were compared with LMM without correlation structure (i.e., the random intercept model) using a likelihood ratio test. Details of these correlation structures are fully described in Pinheiro, Bates [71].

**LMM with variance structure**

To account for variation (i.e., heteroscedasticity due to cluster) not accounted for by the random effects, we also re-fitted the LMM (i.e., the random intercept model) assuming that the residuals were independent on cluster level. In this case, we used the *varIdent* variance function implemented in the *nlme* package [71]. The model was fitted using ML, and compared with LMM (i.e., the random intercept model) using the likelihood ratio test to determine the effect of cluster information on the model accuracy. Finally, the best model as indicated by the likelihood ratio test was refitted using restricted maximum likelihood (REML). To ensure our modelling strategy has accounted for heteroscedasticity due to cluster structure, the residuals from best model was further analyzed by fitting a residual intercept model (i.e., null model) and a residual random intercept model. The two models were compared using a likelihood ratio test to determine if we still have an effect of cluster structure in the residuals. Pseudo R-square ($R^2$) computed as the square of the Pearson correlation coefficient between observed and predicted values was used to assess the quality of the model fit.

**Accuracy assessment**

To enable fair comparisons of best LMM and non-parametric imputations (presented below), the prediction error of the best LMM was estimated by using leave-one-cluster-out cross validation (LOCOCV) [72]. Owing to the number of clusters used in the current study and the lack of an independent validation dataset, LOCOCV was therefore applied. The predicted values of AGB obtained from the LOCOCV $\left(\text{i.e.}, \widehat{SQRT}(AGB)\right)$ were corrected for bias (caused by the square root transformation) using the method by [73] according to

$$\widehat{AGB}_{corrected} = \left(\widehat{SQRT}(AGB)\right)^2 + MSE \tag{4}$$

where $MSE$ is the mean square error of the model. Relative root mean square error from the LOCOCV ($RMSE_{CV}\%$) was used as a criterion for assessing model accuracy and calculated as

$$RMSE_{CV}\% = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}}{\bar{y}} \times 100 \tag{5}$$

where $y_i$ and $\hat{y}_i$ denote ground reference AGB and predicted AGB for plot $i$, respectively, and $\bar{y}$ denotes mean ground reference AGB for all plots. $RMSE_{CV}\%$ is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction [74].

### Non-parametric method

#### $k$-NN imputation

Imputation using $k$-NN is a non-parametric method that has often been used for predicting various attributes in forest inventories supported by remotely sensed auxiliary information [ e.g. 75, 76]. In $k$-NN terminology it is typically distinguished between *reference* and *target* datasets. The population units for which observations of both response and explanatory variables are available is labeled reference set; the set of the population units for with only the explanatory variables are available is termed as the target set. In our study, the reference set contained both ground reference AGB and the ALS metrics, while the target set contained only the ALS metrics.

The similarity between the $i$th target observation and $j$th reference observation was quantified by the means of the Euclidian distances calculated in the feature space as:

$$d_{ij} = \sqrt{(x_i - x_j)'(x_i - x_j)} \tag{6}$$

where $x_i$ and $x_j$ are the feature vectors. Hence, the similarity between the target and reference observations will increase as the $d_{ij}$ distances decreases, and consequently the nearest neighbor of the $i$th target observation is the reference observation located at the shortest Euclidian distance in the feature space.

The imputed value $\hat{y}_i$ is expressed as a weighted sums of the responses taken from the nearest $k$ reference observations as follows:

$$\hat{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i \tag{7}$$

where $y_{ij}^{t}, j = 1,2, \ldots \ldots k$ is the set of the response variable observations for the $k$ reference set elements that are nearest to the *ith* target set elements in the feature space. The $k$-weights associated with the response in equation 6 were obtained as

$$w_{ij} = d_{ij}\left[\sum_{j=1}^{k} d_{ij}\right]^{-1} \quad (8)$$

In order to reduce the data redundancy and improve the overall interpretability, a variable selection procedure was applied by using *varSelection* function in *yaImpute* package [77] of the R software [67]. Model fitting was done by using *knnreg* function in *caret* package [78]. For $k$-NN imputations, selection of $k$ has an influence on the accuracy of the imputation. Large values of $k$ are not recommended since this will shift the predictions towards the sample mean [26]. For this study we used $k$-values of 5 and 10 based on our initial results from cross validation (not presented here). LOCOCV was used to assess the accuracy of k-*NN* imputations, where one cluster at time was used as the target set while the remaining clusters were used at the reference set. To assess the ability of the $k$-NN to account for the dependence and heteroscedasticity due to cluster structure, we computed the residuals from the LOCOCV, then we fitted a residual intercept model and compared with residual random intercept model using likelihood ratio test. Lastly, we compared $k$-NN and LMM using $RMSE_{CV}\%$.

### Assessing the effect of post-stratification on prediction accuracy

To account for the variation in prediction accuracy that might be attributed to the differences in vegetation and land use types, the best LMM (i.e., LMM with variance structure) was further evaluated for different vegetation and land use types. Both relative root mean square errors from model predictions (RMSE%) and LOCOCV ($RMSE_{CV}\%$) were calculated and presented for each category of vegetation and land use type. Specific LMMs (i.e., random intercept models) were fitted for the post-strata as defined by vegetation and land use types.

The models were evaluated using LOCOCV. For each of the post-stratum model. $R^2$ and $RMSE_{CV}$% were computed and compared with the $RMSE_{CV}$% obtained when evaluating the non-post-stratified model for the respective post-stratum.

**Author's contributions**

EWM and LTE have been involved in designing the study, drafting the manuscript, data analysis and write up. OMB has been involved in data analysis and revising the manuscript. EN and TG have been responsible for designing the ALS acquisition and they were involved in revising the manuscript. REM and EZ were involved on planning the field inventory and revising the manuscript. All the authors have read and approved the final manuscript.

**Author's information**

EWM is a PhD student in forest inventory at Norwegian University of Life Sciences (NBMU). He is associated with the forest mensuration group in Department of Ecology and Natural Resources Management at NMBU. ETH and OMB are researchers in the same group specialized on the application of ALS in forestry. EN and TG are senior scientists and professors in forest inventory and forest management at NMBU. Both EN and TG are resource persons for the forest mensuration group at NMBU. REM is professor in forest inventory at Sokoine University of Agriculture (SUA). Tanzania, while EZ is senior lecturer in forest inventory and MRV expert at SUA.

**Acknowledgements**

The financial support for this research was provided by the projects entitled "Enhancing the Measuring. Reporting and Verification (MRV) of forests in Tanzania through the application of advanced remote sensing techniques". The main author acknowledges also the project entitled "Climate Change Impacts. Adaptation and Mitigation (CCIAM) in Tanzania" for financing his study which resulted into this publication. We are grateful to our field team in Tanzania for the field data collection and Terratec, Norway, for collecting and processing of the ALS data. We would also like to acknowledge the administration of Liwale district council for all support, and especially for provision of office space.

**List of abbreviations**

ABA: Area-Based Approach

AGB: Aboveground Biomass

AIC: Akaike Information Criterion

ALS: Airborne Laser Scanning

GLONASS: Global Navigation Satellite System

GPS: Global Positioning System

*k*-NN: *k*- Nearest Neighbors

LMM: Linear Mixed Effects Model

LMMs: Linear Mixed Effects Models

LOCOCV: Leave-One-Cluster-Out Cross Validation

MRV: Measuring, Reporting and Verification

MSE: Mean Square Error

OLS: Ordinary Least Square

REDD+: Reducing Emissions from Deforestation and Forest Degradation

RMSE: Root Mean Square Error

$RMSE_{cv}$: Root Mean Square Error from Cross Validation

SAR: Synthetic Aperture Radar

SE: Standard Error

SSE: Sum of Square Error

TIN: Triangular Irregular Network

UNFCCC: United Nations Framework Convention on Climate Change

### Competing interests

The authors declare that they have no competing interests

### References

1. Martin AR, Thomas SC. A reassessment of carbon content in tropical trees. PLoS One. 2011;6:e 23533.
2. Herold M, Skutsch M. Monitoring, reporting and verification for national REDD plus programmes: two proposals. Environmental Research Letters. 2011;6. doi:10.1088/1748-9326/6/1/014002.
3. Joseph S, Herold M, Sunderlin WD, Verchot LV. REDD+ readiness: early insights on monitoring, reporting and verification systems of project developers. Environmental Research Letters. 2013;8:034038.
4. MNRT. 2015. National Forest Resources Monitoring and Assessment of Tanzania Mainland (NAFORMA). Main Results. http://www.fao.org/forestry/43612-09cf2f02e20b55c1c00569e679197dcdc.pdf. Accessed 17th Aug 2015.
5. Burgess ND, Bahane B, Clairs T, Danielsen F, Dalsgaard S, Funder M et al. Getting ready for REDD+ in Tanzania: a case study of progress and challenges. Oryx. 2010;44:339-51.
6. Tomppo E, Malimbwi R, Katila M, Mäkisara K, Henttonen H, Chamuya N et al. A Sampling design for a large area forest inventory-case Tanzania. Canadian Journal of Forest Research. 2014.
7. Kaasalainen S, Holopainen M, Karjalainen M, Vastaranta M, Kankare V, Karila K et al. Combining Lidar and Synthetic Aperture Radar Data to Estimate Forest Biomass: Status and Prospects. Forests. 2015;6:252-70.
8. Næsset E. Area-Based Inventory in Norway–From Innovation to an Operational Reality. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning - concepts and case studies. Dordrecht, Netherlands: Springer; 2014. p. 215–240.

9. Hansen EH, Gobakken T, Bollandsås OM, Zahabu E, Næsset E. Modeling Aboveground Biomass in Dense Tropical Submontane Rainforest Using Airborne Laser Scanner Data. Remote Sensing. 2015;7:788-807.

10. Ioki K, Tsuyuki S, Hirata Y, Phua M-H, Wong WVC, Ling Z-Y et al. Estimating above-ground biomass of tropical rainforest of different degradation levels in Northern Borneo using airborne LiDAR. Forest Ecology and Management. 2014;328:335-41. doi: http://dx.doi.org/10.1016/j.foreco.2014.06.003.

11. Vauhkonen J, Maltamo M, McRoberts RE, Næsset E. Introduction to Forestry Applications of Airborne Laser Scanning. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer;. p. 1-16.

12. Wulder MA, White JC, Nelson RF, Næsset E, Orka HO, Coops NC et al. Lidar sampling for large-area forest characterization: A review. Remote Sensing of Environment. 2012;121:196-209. doi:http://dx.doi.org/10.1016/j.rse.2012.02.001.

13. Gobakken T, Næsset E, Nelson R, Bollandsås OM, Gregoire TG, Ståhl G et al. Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment. 2012;123:443-56. doi:http://dx.doi.org/10.1016/j.rse.2012.01.025.

14. Gregoire TG, Ståhl G, Næsset E, Gobakken T, Nelson R, Holm S. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research. 2010;41:83-95.

15. McRoberts RE, Andersen H-E, Næsset E. Using Airborne Laser Scanning Data to Support Forest Sample Surveys. Forestry Applications of Airborne Laser Scanning. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer; 2014. p. 269-92.

16. Fassnacht FE, Hartig F, Latifi H, Berger C, Hernández J, Corvalán P et al. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sensing of Environment. 2014;154:102-14. doi:http://dx.doi.org/10.1016/j.rse.2014.07.028.

17. Garcia-Gutierrez J, Gonzalez-Ferreiro E, Riquelme-Santos JC, Miranda D, Dieguez-Aranda U, Navarro-Cerrillo RM. Evolutionary feature selection to estimate forest stand variables using LiDAR. International Journal of Applied Earth Observation and Geoinformation. 2014;26:119-31. doi:http://dx.doi.org/10.1016/j.jag.2013.06.005.

18. Garcıa-Gutiérreza J, Martınez-Álvarezb F, Troncosob A, Riquelmea J. A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables.

19. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2012.

20. García S, Fernández A, Luengo J, Herrera F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences. 2010;180:2044-64. doi:http://dx.doi.org/10.1016/j.ins.2009.12.010.

21. Fox JC, Ades PK, Bi H. Stochastic structure and individual-tree growth models. Forest Ecology and Management. 2001;154:261-76. doi:http://dx.doi.org/10.1016/S0378-1127 00632-0.

22. Tang M, Slud EV, Pfeiffer RM. Goodness of fit tests for linear mixed models. Journal of Multivariate Analysis. 2014;130:176-93. doi:http://dx.doi.org/10.1016/j.jmva.2014.03.012.

23. Gałecki A, Burzykowski T. Linear Mixed-effects Models Using R: A Step-by-step Approach. Springer; 2013.

24. Salas C, Ene L, Gregoire TG, Næsset E, Gobakken T. Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models.

Remote Sensing of Environment. 2010;114:1277-85. doi:http://dx.doi.org/10.1016/j.rse.2010.01.020.

25. Packalén P, Maltamo M. The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment. 2007;109:328-41. doi:http://dx.doi.org/10.1016/j.rse.2007.01.005.

26. Hudak AT, Crookston NL, Evans JS, Hall DE, Falkowski MJ. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment. 2008;112:2232-45. doi:http://dx.doi.org/10.1016/j.rse.2007.10.009.

27. Eskelson BN, Temesgen H, Lemay V, Barrett TM, Crookston NL, Hudak AT. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research. 2009;24:235-46.

28. Ene LT, Næsset E, Gobakken T, Gregoire TG, Ståhl G, Holm S. A simulation approach for accuracy assessment of two-phase post-stratified estimation in large-area LiDAR biomass surveys. Remote Sensing of Environment. 2013;133:210-24. doi:http://dx.doi.org/10.1016/j.rse.2013.02.002.

29. McRoberts RE, Tomppo EO, Finley AO, Heikkinen J. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sensing of Environment. 2007;111:466-80.

30. Beaudoin A, Bernier P, Guindon L, Villemaire P, Guo X, Stinson G et al. Mapping attributes of Canada's forests at moderate resolution through k NN and MODIS imagery. Canadian Journal of Forest Research. 2014;44:521-32.

31. Chirici G, Corona P, Marchetti M, Mastronardi A, Maselli F, Bottai L et al. k-NN FOREST: a software for the non-parametric prediction and mapping of environmental variables by the k-Nearest Neighbors algorithm. 2012.

32. Penner M, Pitt D, Woods M. Parametric vs. nonparametric LiDAR models for operational forest inventory in boreal Ontario. Canadian Journal of Remote Sensing. 2013;39:426-43.

33. Gagliasso D, Hummel S, Temesgen H. A comparison of selected parametric and non-parametric imputation methods for estimating forest biomass and basal area. Open Journal of Forestry. 2014;4:42.

34. Bollandsås OM, Maltamo M, Gobakken T, Næsset E. Comparing parametric and non-parametric modelling of diameter distributions on independent data using airborne laser scanning in a boreal conifer forest. Forestry. 2013;86:493-501.

35. Leitold V, Keller M, Morton DC, Cook BD, Shimabukuro YE. Airborne lidar-based estimates of tropical forest structure in complex terrain: opportunities and trade-offs for REDD+. Carbon balance and management. 2015;10:3.

36. Latifi H, Fassnacht FE, Hartig F, Berger C, Hernández J, Corvalán P et al. Stratified aboveground forest biomass estimation by remote sensing data. International Journal of Applied Earth Observation and Geoinformation. 2015;38:229-41. doi:http://dx.doi.org/10.1016/j.jag.2015.01.016.

37. Næsset E, Gobakken T. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sensing of Environment. 2008;112:3079-90. doi:http://dx.doi.org/10.1016/j.rse.2008.03.004.

38. Næsset E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment. 2002;80:88-99. doi:http://dx.doi.org/10.1016/S0034-4257(01)00290-5.

39. Huang W, Sun G, Dubayah R, Cook B, Montesano P, Ni W et al. Mapping biomass change after forest disturbance: Applying LiDAR footprint-derived models at key map scales. Remote Sensing of Environment. 2013;134:319-32. doi:http://dx.doi.org/10.1016/j.rse.2013.03.017.

40. Chen Q, Vaglio Laurin G, Battles JJ, Saah D. Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. Remote Sensing of Environment. 2012;121:108-17. doi:http://dx.doi.org/10.1016/j.rse.2012.01.021.

41. Dewees PA, Campbell BM, Katerere Y, Sitoe A, Cunningham AB, Angelsen A et al. Managing the Miombo woodlands of southern Africa: policies, incentives and options for the rural poor. Journal of natural resources policy research. 2010;2:57-73.

42. Ribeiro NS, Matos CN, Moura IR, Washington-Allen RA, Ribeiro AI. Monitoring vegetation dynamics and carbon stock density in miombo woodlands. Carbon balance and management. 2013;8:1-9.

43. Haara A, Kangas A. Comparing k Nearest Neighbours Methods and Linear Regression–Is There Reason To Select One Over the Other? Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS). 2012;4:Pages: 50-65.

44. Fehrmann L, Lehtonen A, Kleinn C, Tomppo E. Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. Canadian journal of forest research. 2008;38:1-9.

45. Baffetta F, Corona P, Fattorini L. Design-based diagnostics for k-NN estimators of forest resources. Canadian journal of forest research. 2010;41:59-72.

46. Baffetta F, Fattorini L, Franceschi S, Corona P. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment. 2009;113:463-75.

47. Ståhl G, Holm S, Gregoire TG, Gobakken T, Næsset E, Nelson R. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Canadian journal of forest research. 2010;41:96-107.

48. Mascaro J, Asner GP, Dent DH, DeWalt SJ, Denslow JS. Scale-dependence of aboveground carbon accumulation in secondary forests of Panama: A test of the intermediate peak hypothesis. Forest Ecology and Management. 2012;276:62-70.

49. Asner GP, Mascaro J, Muller-Landau HC, Vieilledent G, Vaudry R, Rasamoelina M et al. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia. 2012;168:1147-60.

50. Asner GP. Tropical forest carbon assessment: integrating satellite and airborne mapping approaches. Environmental Research Letters. 2009;4:034009.

51. Asner GP, Powell GV, Mascaro J, Knapp DE, Clark JK, Jacobson J et al. High-resolution forest carbon stocks and emissions in the Amazon. Proceedings of the National Academy of Sciences. 2010;107:16738-42.

52. Vaglio Laurin G, Chen Q, Lindsell JA, Coomes DA, Frate FD, Guerriero L et al. Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. ISPRS Journal of Photogrammetry and Remote Sensing. 2014;89:49-58. doi:http://dx.doi.org/10.1016/j.isprsjprs.2014.01.001.

53. Clark ML, Roberts DA, Ewel JJ, Clark DB. Estimation of tropical rain forest aboveground biomass with small-footprint lidar and hyperspectral sensors. Remote Sensing of Environment. 2011;115:2931-42. doi:http://dx.doi.org/10.1016/j.rse.2010.08.029.

54. Mascaro J, Detto M, Asner GP, Muller-Landau HC. Evaluating uncertainty in mapping forest carbon with airborne LiDAR. Remote Sensing of Environment. 2011;115:3770-4.

55. Mauya E, Hansen E, Gobakken T, Bollandsås O, Malimbwi R, Næsset E. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. Carbon Balance and Management. 2015;10:1-14. doi:10.1186/s13021-015-0021-x.

56. Zolkos S. Goetz S. Dubayah R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment. 2013;128:289-98.

57. Frazer GW, Magnussen S, Wulder MA, Niemann KO. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sensing of Environment. 2011;115:636-49. doi:10.1016/j.rse.2010.10.008.

58. Mugasha WA, Eid T, Bollandsås OM, Malimbwi RE, Chamshama SAO, Zahabu E et al. Allometric models for prediction of above- and belowground biomass of trees in the miombo woodlands of Tanzania. Forest Ecology and Management. 2013;310:87-101. doi:http://dx.doi.org/10.1016/j.foreco.2013.08.003.

59. LDC. Liwale Distric Socio-Economic Profile. Unpublished Report 2014. p. 56.

60. Cochran W. G.(1977): Sampling techniques. New York. Wiley and Sons. 1977;98:259-61.

61. Anon. Pinnacle user's manual. Javad Positioning Systems. CA. Edited by Jose S. USA.1999.

62. MNRT. NAFORMA Field Manual – Biophysical. http:// http://www.fao.org/forestry/23484-05b4a328815ecc769685b21b03be44ea77.pdf (2011). Accessed 23rd Feb 2014

63. Axelsson P. Processing of laser scanner data—algorithms and applications. ISPRS Journal of Photogrammetry and Remote Sensing. 1999;54:138-47.

64. Axelsson P. DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing. 2000;33:111-8.

65. Næsset E. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research. 2004;19:164-79.

66. Lumley T. Package 'LEAPS': Regression subset selection. R package, version. 2012;2.

67. Team RC. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. Open access available at: http://cran r-project org. 2014.

68. Tsui OW, Coops NC, Wulder MA, Marshall PL, McCardle A. Using multi-frequency radar and discrete-return LiDAR measurements to estimate above-ground biomass and biomass components in a coastal temperate forest. ISPRS Journal of Photogrammetry and Remote Sensing. 2012;69:121-33.

69. Mallows CL. Some comments on C p. Technometrics. 1973;15:661-75.

70. Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. Springer; 2009.

71. Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. Springer; 2000.

72. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

73. Gregoire TG, Lin QF, Boudreau J, Nelson R. Regression estimation following the square-root transformation of the response. Forest Science. 2008;54:597-606.

74. Yoo S, Im J, Wagner JE. Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. Landscape and Urban Planning. 2012;107:293-306. doi:http://dx.doi.org/10.1016/j.landurbplan.2012.06.009.

75. McRoberts RE. Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management. 2012;272:3-12.

76. McRoberts RE, Tomppo EO. Remote sensing support for national forest inventories. Remote Sensing of Environment. 2007;110:412-9.

77. Crookston NL, Finley AO. yaimpute: An r package for knn imputation. Journal of Statistical Software. 2008;23:1-16.

78. Engelhardt A, Kuhn MM. Package 'caret'. 2009.

# PAPER III

Mauya et al. Carbon Balance and Management (2015) 10:10
DOI 10.1186/s13021-015-0021-x

Carbon Balance and Management
a SpringerOpen Journal

RESEARCH Open Access

# Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania

Ernest William Mauya[1*], Endre Hofstad Hansen[1], Terje Gobakken[1], Ole Martin Bollandsås[1], Rogers Ernest Malimbwi[2] and Erik Næsset[1]

**Abstract**

**Background:** Airborne laser scanning (ALS) has recently emerged as a promising tool to acquire auxiliary information for improving aboveground biomass (AGB) estimation in sample-based forest inventories. Under design-based and model-assisted inferential frameworks, the estimation relies on a model that relates the auxiliary ALS metrics to AGB estimated on ground plots. The size of the field plots has been identified as one source of model uncertainty because of the so-called boundary effects which increases with decreasing plot size. Recent research in tropical forests has aimed to quantify the boundary effects on model prediction accuracy, but evidence of the consequences for the final AGB estimates is lacking. In this study we analyzed the effect of field plot size on model prediction accuracy and its implication when used in a model-assisted inferential framework.

**Results:** The results showed that the prediction accuracy of the model improved as the plot size increased. The adjusted $R^2$ increased from 0.35 to 0.74 while the relative root mean square error decreased from 63.6 to 29.2%. Indicators of boundary effects were identified and confirmed to have significant effects on the model residuals. Variance estimates of model-assisted mean AGB relative to corresponding variance estimates of pure field-based AGB, decreased with increasing plot size in the range from 200 to 3000 $m^2$. The variance ratio of field-based estimates relative to model-assisted variance ranged from 1.7 to 7.7.

**Conclusions:** This study showed that the relative improvement in precision of AGB estimation when increasing field-plot size, was greater for an ALS-assisted inventory compared to that of a pure field-based inventory.

**Keywords:** Airborne laser scanning, Model-assisted estimation, Plot size, Aboveground biomass

## Background

Tropical forests play an important role in the global carbon cycle as they store about 40% of the global terrestrial carbon, and absorb larger amounts of $CO_2$ from the atmosphere than any other vegetation type [1]. Despite their potential, tropical forests continue to be exploited at alarming rates, by being converted into secondary forest and many other forms of land use. In an effort to conserve tropical forests, the United Nations Framework Convention on Climate Change (UNFCCC) has developed the mechanism called Reducing Emissions from Deforestation and Forest Degradation in tropical countries (REDD+). There is high interest in seeing such initiatives to take form, but a key limitation for successful implementation of REDD+ is reliable methods for quantifying forest aboveground biomass (AGB) [2,3]. Such methods are important because payments for carbon offsets under REDD+ are based on estimates of carbon stock and stock changes over time. Moreover, AGB information is also useful for understanding the contribution of the tropical forests to the global carbon cycle and ecosystem processes [4].

* Correspondence: ernest.mauya@nmbu.no
[1]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, Oslo, NO 1432, Ås, Norway
Full list of author information is available at the end of the article

Airborne laser scanning (ALS) has emerged as one of the most promising remote sensing technologies to support AGB forest inventories in boreal-, temperate-, and tropical forests [5]. A particular strength of ALS for forest applications is its ability to accurately characterize the three-dimensional (3D) structure of the forest canopy [6]. Such information is more useful for forest inventories than the information from other remote sensing techniques see e.g. [7]. Height and density metrics derived from the ALS data has been reported to be highly correlated with AGB see e.g. [8,9]. Furthermore, ALS has shown to be superior to other remote sensing data sources because the relationship between AGB and the remotely sensed information has a much higher saturation level for ALS compared to other types remote sensing. Because of this, ALS is a highly appropriate choice of technique in high-biomass forests. Based on its potential, ALS has recently been recommended for Monitoring, Reporting and Verification (MRV) systems under REDD+ initiatives [10].

Estimation of AGB using ALS is often carried out according to the area-based approach (ABA) [11]. In ABA, empirical models between various metrics derived from the ALS data and AGB values obtained in georeferenced field sample plots are fitted. The area of interest is then tessellated into grid cells [12] with the same size as the plots [13,14] and the developed models are used to provide cell-wise predictions of AGB. Finally, estimates for the particular area of interest (forest stand, forest property, village, district, or nation) are provided by summing the individual cell predictions. For some estimation approaches, adjustment of model prediction bias [15] is also carried out.

As indicated above, the modeled relationship between ALS metrics and ground-based values is of fundamental importance for the outcome of the ALS-assisted estimation. The use of field plot data for model development requires co-registration of field plot location with the ALS data [16,17]. In an ALS-assisted inventory, the point cloud is extracted only within the plot perimeter. However, in field measurements trees are treated as being inside plots if the center point of the stem is inside the plot. This is a challenge in ALS-assisted forest inventory, since the crowns of trees just outside the plot border partly extend into the plot area which means that the ALS data will be affected by trees that are not registered in field. Conversely, also trees just inside the plot extend their crowns beyond the plot boundary. This means that there may be mismatch between the data captured in field and from the air.

In order to reduce these boundary effects, it has been suggested in a number of studies to use larger plots in ALS-assisted forest inventory see e.g. [18,19]. This is because, as plot size increases, the perimeter to area ratio decreases and thus the plots include a lower proportion of boundary-related elements. Similarly, the relative and negative influence of a given plot positioning error is reduced because the relative overlap between the field- and ALS-data becomes larger as plot size increases. Reduction in model errors are also expected by increasing plot size due to so-called spatial averaging of the errors [20], because both the field observations and the ALS data capture more of the spatial variation as they increase in size. Thus, as plot sizes increase, the variances of field-based and ALS-assisted estimates are expected to be reduced, which means that fewer plots are needed to reach a certain precision of an AGB estimate. However, large plots also have disadvantages by being more complicated to measure, which may affect the time consumption for collecting field measurements [21]. This makes it challenging to select the "optimal" plot size that balances the tradeoff between plot size, sample size (number of plots), on-plot costs, traveling costs and precision of ALS-assisted AGB estimates in different forest types.

As indicated above, plot size has a profound effect on the precision of ALS-assisted AGB estimates for several reasons. Likewise, the plot size has an impact on the precision of pure field-based estimates for reasons mentioned above: larger plots capture more of the variability in the area of interest and thus precision will tend to improve as long as the sample size is kept constant. A key question is therefore if larger plots will favor ALS-assisted estimation precision to the same extent as it favors field-based estimation precision. Different responses to plot size should have a direct impact on how tropical ALS-assisted field sample surveys should be designed as their designs currently are "optimized" for pure field-based estimation.

Forest sample surveys are often designed according to design-based (probability-based) principles. Simple random sampling is one of these principles, and analytical and so-called design-unbiased estimators and corresponding variance estimators exist for a great number of such designs. When auxiliary data such as those acquired by ALS are at hand for the entire area of interest, or at least with partial coverage of the area of interest, use of these data can greatly improve the precision over a pure field-based estimate assuming the same design. The inferential framework applied under probability sampling when a model is used to predict AGB using the ALS data is known as design-based model-assisted (MA) estimation. In the MA framework, the model is used to predict AGB for grid cells and then AGB is summed over all grid cells as indicated in the ABA, but in addition to that, the model predictions for the ground samples are used to provide an estimate of bias in the model predictions, which corrects the pure model-based estimate. Several studies see e.g. [22-24] have indicated the potential of MA estimation in reducing the variance of AGB estimates in boreal forests, but apart from some indications provided by [23], neither of them has analyzed how the variance of the estimates is affected by changes in field plot sizes. In tropical forests where the current study was conducted, there is even less knowledge regarding performance of MA estimation using ALS with varying plot sizes. Several tropical studies have examined the effects of plot size on model prediction accuracy See e.g. [25-27], but none of them have assessed the effects on the precision of AGB estimates and compared such precision estimates with corresponding precision of field-based AGB estimates using the same sampling design, which is of fundamental importance for designing future sample surveys serving multiple purposes and estimation approaches.

The objectives of this study were to (1) examine the effects of field plot size on AGB regression model quality, (2) assess plot boundary effect and its impact on model quality based on the field data, and (3) quantify the precision of ALS-assisted estimates of AGB relative to field-based estimates of AGB assuming the same design for different plot sizes. The study was conducted in tropical rain forest in Tanzania with high AGB densities, which was expected to represent a particular challenge in terms of large boundary effects.

## Results

### Effects of field plot size on ALS AGB predictions

To assess the effect of plot size on ALS assisted forest inventory, we first fitted the regression models for each of the plot sizes. The independent variables selected varied between the models developed for the different plot sizes (Table 1). The number of variables varied between two and three. For all models, the parameter estimates were significantly different from zero ($p < 0.05$) and the VIF values were <10, indicating acceptable levels of multicolinearity. The variability explained by separate models (i.e adjusted $R^2$) improved as the plot size increased, with few exceptions (Figure 1a). The adjusted $R^2$ ranged from 0.35 for the plot size of 200 $m^2$ to 0.74 for the plot size of 3000 $m^2$. The RMSE% values for LOOCV decreased non-linearly with increasing plot size, from 63.8 to 29.2% (Figure 1b). The MPE% values (Figure 1b) and the pattern of under predictions for plots with high AGB were relatively lower for larger plots compared smaller (Figure 2). However, it should be noted that the number of the larger plots was relatively small.

### Boundary effects

Boundary effects were studied by analyzing how the relative residual errors of the models were affected by the ground reference AGB of the trees in an outer buffer zone for different field plot sizes. Our results showed that $SAGB_{buffer}$ and $MAGB_{buffer}$ contributed to explaining the variation in the relative residual errors (Table 2). Relating the absolute value of the relative residual with plot size using simple linear regression model indicated that there was a highly significant effect of plot size ($p < 0.0001$). Furthermore, the parameter estimate for plot size was negative showing that the relative residual is larger in absolute terms for small plots compared to larger plots (Table 3).

**Table 1 Selected ALS metrics for different plot sizes**

| Plot size ($m^2$) | n | Selected variables[a] | | |
|---|---|---|---|---|
| 200 | 30 | D F | D L | log H F |
| 300 | 30 | D L | log H F | log D L |
| 400 | 30 | H F | D L | |
| 500 | 30 | H F | D L | |
| 600 | 30 | H F | D L | |
| 700 | 30 | H F | D L | |
| 800 | 30 | H F | D L | |
| 900 | 30 | H F | D L | |
| 1000 | 30 | Hsd L | D L | log D F |
| 1100 | 30 | $D_1F$ | $D_1L$ | log H F |
| 1200 | 30 | $H_{mean}F$ | D L | |
| 1300 | 30 | H F | D L | |
| 1400 | 30 | $D_3F$ | $D_2L$ | log H F |
| 1500 | 30 | H F | D L | |
| 1600 | 30 | H F | D L | |
| 1700 | 30 | H F | D L | |
| 1800 | 30 | H F | D L | |
| 1900 | 30 | H F | D L | |
| 2000 | 25 | H F | D L | |
| 2100 | 25 | H F | D L | |
| 2200 | 24 | H F | D L | |
| 2300 | 24 | H F | D L | |
| 2400 | 24 | H F | D F | |
| 2500 | 22 | H F | D F | |
| 2600 | 22 | H F | D F | |
| 2700 | 22 | H F | D F | |
| 2800 | 22 | H F | D F | |
| 2900 | 22 | H F | D F | |
| 3000 | 22 | H F | D F | |

[a] $D_0F$, $D_1F$ and $D_3F$ – Canopy densities corresponding to the proportion of first echoes above fraction #0 (2 m), #1 and #3 (see text). [a] $D_0L$, $D_1L$ and $D_2L$ – Canopy densities corresponding to the proportion of last echoes above fraction #0 (2 m), #1 and #2 (see text)

$H_{10}F$, $H_{60}F$, $H_{70}F$, $H_{80}F$ and $H_{90}F$ – ALS height percentiles of the canopy height for the first echo

Hsd L – Standard deviation of the canopy height of the first echoes.

$H_{mean}F$ – Arithmetic mean of the first echo ALS canopy height.

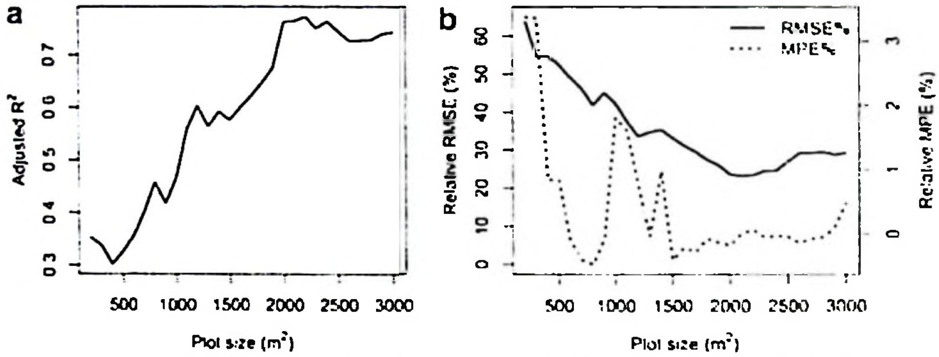



**Figure 1** Model quality and plot size. **(a)** Adjusted $R^2$ versus plot sizes. **(b)** Relative MPE% and RMSE% versus plot sizes.

### Efficiency of ALS-assisted AGB estimation

The SE estimates for the field-based AGB estimates were larger than the corresponding model-assisted SE estimates (Figure 3). For the plot sizes that allowed consistent analysis for all 30 sizes, i.e. from 200 to 1900 $m^2$, the field-based SE estimates decreased from 58.0 Mg $ha^{-1}$ to 28.7 Mg $ha^{-1}$, while the model-assisted SE estimates decreased from 44.3 Mg $ha^{-1}$ to 15.5 Mg $ha^{-1}$. Relative to the mean of field reference AGB for the plot size from 200 to 1900 $m^2$, the field -based SE estimates decreased from 14.1% to 8.2% , while for the model-assisted estimates decreased from 10.8% to 4.4%. Similarly, for the larger plots (up to 3000 $m^2$) for which 22 observations were available for consistent analysis, the SE estimates for model-assisted were relatively much smaller compared to the field-based inventory. In both cases the SE was higher for smaller plots compared to the larger plots. Generally, the effectiveness of the ALS-assisted estimates was more improved as the plot size increased compared to the field-based estimates. This indicates that larger plots are relatively more favorable for ALS-assisted estimation than for pure field-based estimation. The RE values were >1 with a maximum value of 3.4 (Figure 4) for the plot sizes ranging from 200–1900 $m^2$

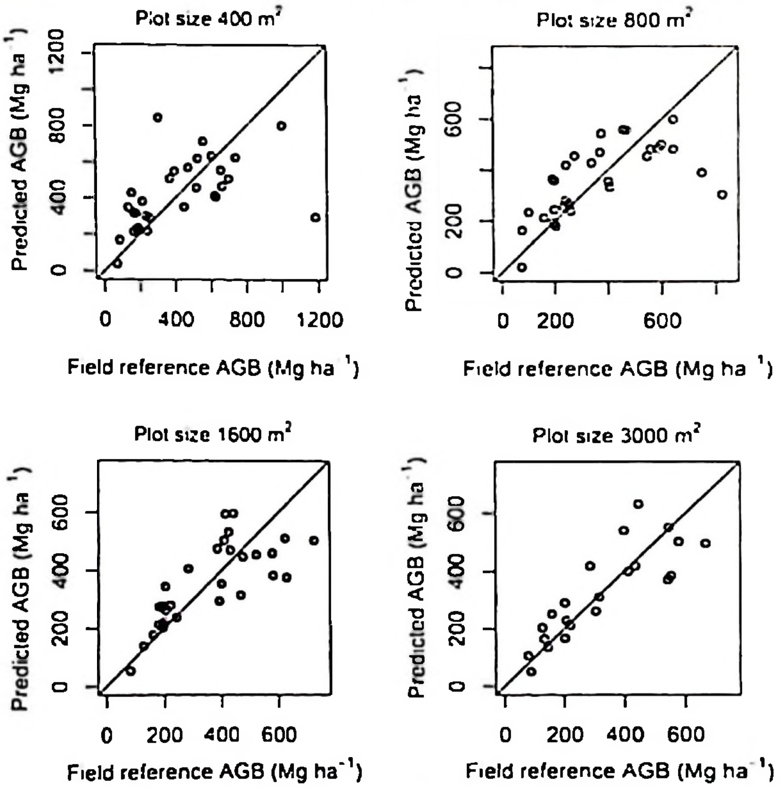



**Figure 2** Relationship between field reference AGB and predicted AGB for different plot sizes.

**Table 2 Coefficient estimates for models explaining residual errors of AGB using information extracted from buffer zones**

| Models[1] | Model parameter | 3 m buffer | | | 6 m buffer | | |
|---|---|---|---|---|---|---|---|
| | | Parameter estimate | p-value | AIC | Parameter estimate | p-value | AIC |
| Model 1 | Intercept | 0.0156 | 0.8022 | 297 | 0.1835 | 0.0206 | 654 |
| | [2]$SAGB_{buffer}$ | 0.0548 | <0.0001 | | 0.1892 | <0.0001 | |
| Model 2 | Intercept | 0.0321 | 0.5826 | 266 | 0.0501 | 0.4674 | 663 |
| | [2]$MAGB_{buffer}$ | 0.4865 | <0.0001 | | 0.7015 | <0.0001 | |

[1]Models = Two models. Model 1 uses $SAGB_{buffer}$ as fixed effects with plot identity as random effect. Model 2 uses $MAGB_{buffer}$ with plot identity as random effect (see text).

[2]$SAGB_{buffer}$ = Ratio of either sum of AGB at the buffer to the ground reference AGB per hectare, $MAGB_{buffer}$ = ratio of Maximum AGB at the buffer to the ground reference AGB per hectare (see text).

for which we have a complete dataset of 30 plots. For the other set with plot size up to 3000 $m^2$ the maximum RE value was 7.7. It should be noted that the peak in relative efficiency for the smallest dataset (22 plots) in Figure 4 was caused by considerable change in the observed AGB for a single plot when increasing the plot size beyond 2000 $m^2$. The increasing AGB was due to a large tree that was included in the plot measurements once the plot radius exceeded 25 m. This illustrates that in a small dataset the results can be sensitive to individual observations and even to the presence of individual trees.

## Discussion

The findings of this study demonstrated the importance of choosing appropriate field plot sizes in ALS-assisted forest inventories in tropical forests. This is particularly important given that field campaigns are expensive and time consuming, and linking field measurements with remotely sensed data in the most effective manner would benefit both REDD+ implementations, together with all other studies related to forest carbon cycle. The current study extends previous research conducted in tropical forests, by having a dataset with a wide range of plot sizes. Furthermore, most of the previous studies have used rectangular plots.

See e.g. [18,26], whereas in this case circular plots have been used. Circular plots are more convenient for remote sensing studies compared to square or rectangular plots because only a single coordinate together with a plot radius are needed to match the two data sources geographically [19,28,29]. Circular plots are also within certain sizes easier to establish in the field because they have one dimension (i.e. radius) that defines the plot boundary. The use of circular plots minimizes the plot boundary effects because of a smaller circumference to area ratio than all other plot shapes. However, the visibility from the plot center to the perimeter on a circular plot is increasingly hampered as the plots get larger, which increase per tree measurement time for the border trees. An increase of the area of a rectangular plot would not necessarily mean increased marginal cost (cost of including one more tree) if the width of the plot is kept constant and inclusion of trees are made with reference to the long side. However, rectangular plots are in general more difficult to establish. For example, in rugged terrain it can be difficult to keep the sides parallel.

**Table 3 Parameter estimates for the model relating relative residual in absolute form and plot sizes**

| Coefficients | Parameter estimates | p-value |
|---|---|---|
| Intercept | 0.5060 | <0.0001 |
| Plot size | 0.0002 | <0.0001 |

Our findings demonstrated empirically the positive effects of increasing plot sizes on improved predictive power of the AGB models. The model fit (adjusted $R^2$) of the regression models was improved as plot size increased. Reduced circumference to area ratio, spatial averaging, and less effect of positioning errors are probably the main reasons. The fit of our models are in line with previous ALS-based studies in both tropical forests and temperate forests. For example, [30] reported $R^2$ of 0.78 in the tropical rainforest of Hawaii islands while [31] reported $R^2$ of 0.64 in a tropical rainforest of West Africa. Furthermore, results from the cross-validation showed smaller RMSE% and MPE% (Figure1b) for larger plots compared to smaller plots. Similar trends have been reported and discussed by other authors in both temperate and tropical forests see e.g. [32].

Plot boundary effects have been discussed in previous studies see e.g. [16,33] as one among the sources of model error in ALS-assisted inventories, particularly when relying on small plots. We demostrated this in two steps; first by relating relative residuals to the sum of AGB per hectare for all trees in the buffer ($SAGB_{buffer}$) and the maximum AGB per hectare for the largest tree in the buffer ($MAGB_{buffer}$) where we noted that their importance were depending on the size of the buffer. The buffer conditions as expressed both by ($MAGB_{buffer}$) and ($SAGB_{buffer}$), seemed to have more impact on the

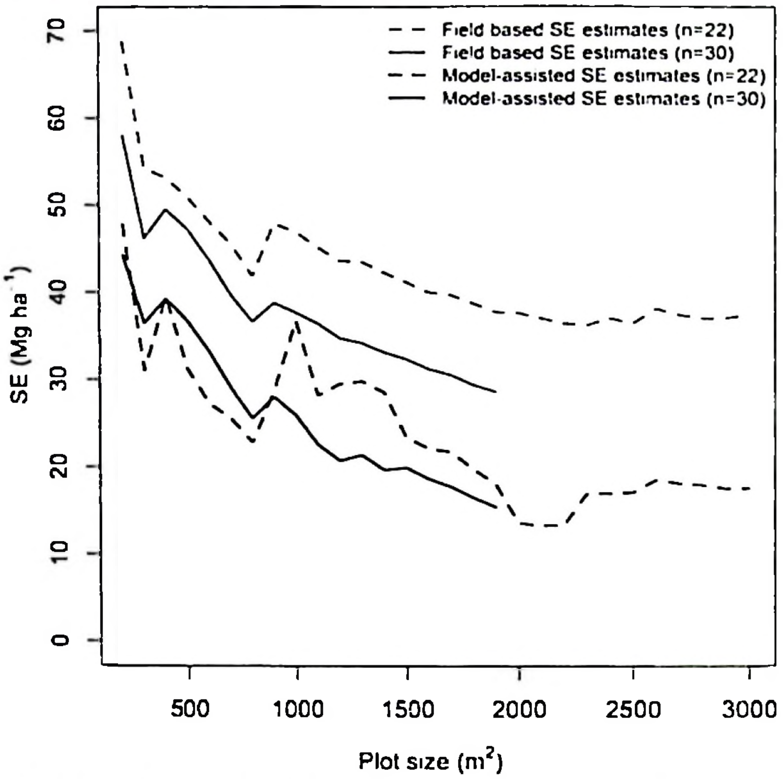



**Figure 3** Field based and model assisted SE estimates for different plot sizes covered in two sample datasets (i.e. 200 to 1900 and 200 to 3000 m²)

residual error with decreasing distance to the plot judged by the AIC values (Table 2), which is logical. Furthermore, when comparing the two variables, $SAGB_{buffer}$ seemed to lose less explanatory power by going from 3 meter to 6 m buffer than $MAGB_{buffer}$. This result was also expected because the represetation of the whole buffer by $SAGB_{buffer}$ is less prone to be changed by the increase in size compared to $MAGB_{buffer}$ which is calculated from a single tree. Furthermore, the decrease in ALS model residuals (Table 3) with increasing plot sizes is a clear indication that smaller plots are more prone to boundary effects compared to larger plots.

Contribution of ALS data in improving precision of AGB estimates was also demonstrated within varying ranges of plot sizes. The RE values were > 1, indicating that ALS-assisted estimation is more efficient compared to pure field-based estimation. To achieve similar precision of a pure field-based estimate relying on simple random sampling, would mean to increase the sample size for the field-based inventory by a factor equivalent to the value of RE, which would have a substantial effect on field inventory costs. In general, the gain in relative efficiency was more pronounced as plot size increased, suggesting that larger plots are more favorable when ALS-data are used to assist in the estimation. Even though we did not undertake any analysis of cost-efficiency, the trend would be toward larger and fewer plots as one introduces ALS to support in the estimation. Even this finding can be attributed to the effects discussed above, namely reduced boundary effects and co-registration errors

Despite the potential of improving the efficiency of ALS-assisted inventories by use of larger plots, choice of an "optimal" plot size must be seen in a broader context by considering a number of factors including; sample sizes, on-plot costs, traveling costs and overall field inventory design. Several authors see e.g. [20,23,30] have indicated that selection of the plot size also will depend on forest types, available resources and the needed precision. Based on our findings, there is larger potential of gaining efficiency of using ALS data in this type of forest when the field plot size is larger than 1200 m². Finally, even though our study was limited to the tropical rainforests of Tanzania, the major findings are of interest and efforts should be taken to upscale to other tropical forests by considering more factors that would lead to selection of "optimal" plot size.

## Conclusions

To conclude, our study has demonstrated that field plot size effect the prediction accuracy of ALS-assisted AGB estimation in the tropical forests. Generally, there was substantial improvement in prediction accuracy from

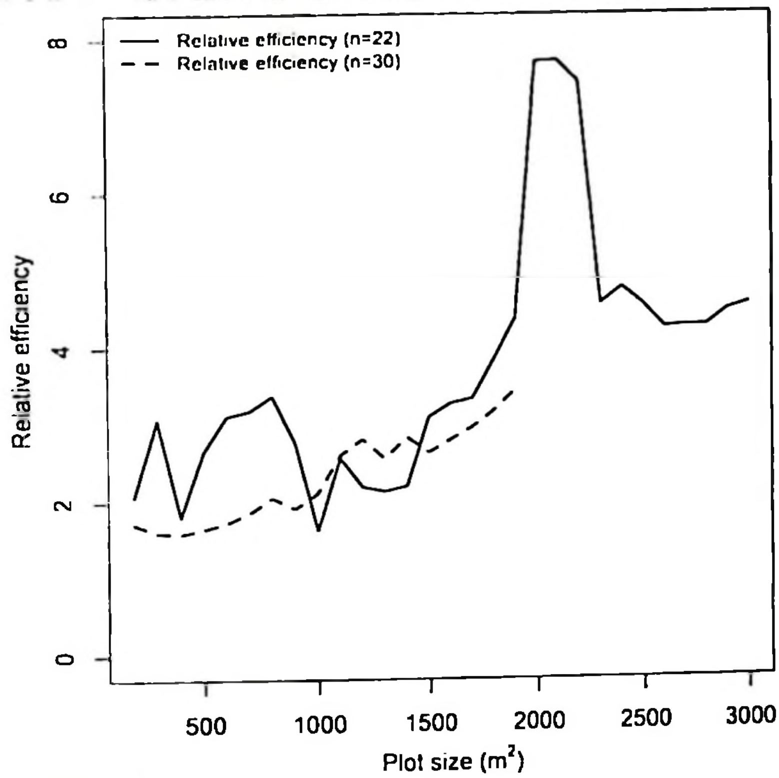



**Figure 4** Relative efficiency for different plot sizes

larger plots compared to smaller plots. Indicators of boundary effects were also identified and confirmed to have significant effects on the model quality. From a purely technical point of view, our results suggested that it is relatively more favorable to increase the plot size when ALS is used to enhance the estimates. This study showed that there is a relative improvement in precision of ALS-assisted AGB estimation, compared to pure field-based estimation up to around 3000 $m^2$ in this type of forest. However, the maximum plot size of 3000 $m^2$ in the current study leaves an open question as to whether there are any additional gains in relative precision beyond this size. Future studies should be conducted to quantify the contribution of ALS to improve estimation precision for even larger plots as the basis for design of future inventories in tropical rainforests. Similar studies should also be conducted in other types of tropical forests.

## Methods

### Site description

The study was conducted in Amani nature reserve (ANR), which is situated in the southern part of the East Usambara Mountains in northern Tanzania (Figure 5). It was gazetted in 1997 with a protected area of 8,380 ha. ANR lies between 5°14' - 5° 04' S and 38° 30' - 38°40' E, with an altitudinal range of 190 to 1130 m above sea level [34]. Rainfall is heavy at higher altitudes and in the southeast of the mountain, with an average of 1900 mm annually. The dry seasons are from June to August and January to March, but rainfall is frequent throughout the year. The mean annual temperature is 20.6°C [35].

### Data collection

#### *Sampling design*

An initial probability sample of 173 field plots with an average size of 900 $m^2$ were established across ANR according to a systematic design (450 m × 900 m distance between plots) in 1999–2000 by a non-governmental conservation and development organization, Frontier Tanzania [34] (Figure 5). The plots were revisited and re-measured in 2008–2012. In order to analyse plot size effects on AGB estimates, a small sub-sample of 30 large plots was established. Measurements on the 30 plots were acquired in a separate campaign after completion of measurements of the large sample. Due to high travel costs and long walking distances in the very steep and rough terrain, establishing a probability sample of 30 large plots across the entire study area was cost-prohibitive. Instead we developed a sampling strategy by which we took advantage of the a priori knowledge of the distribution of AGB in the large probability sample and selected purposefully three sub-regions within the study area in which the initial plots were revisited. There is a strong altitude-dependent AGB gradient in

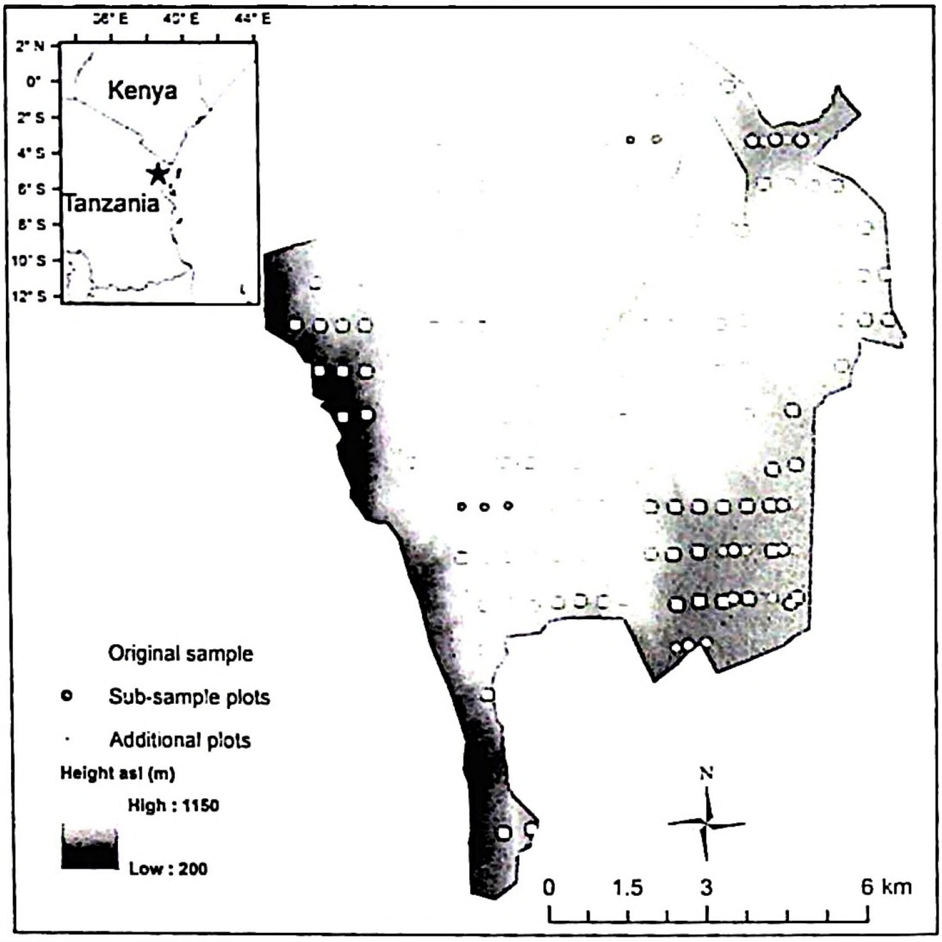



**Figure 5** Study area and field plots layout. Left: Location of Amani nature reserve (marked with star). Right: Map of Amani nature reserve and the two samples of field plots.

the study area. It was therefore important to capture the altitude gradient in each of the three sub-regions in order to resemble the AGB distribution in the initial probability sample.

In the sampled sub-regions, we first selected 16 of the plots in the initial probability sample for measurement. We also established 14 new and additional plots along the grid-lines of the probability sample and located them exactly mid-way between two existing plots. Thus, the distance between our plots was 225 m rather than 450 m.

Although the resulting sample of 30 large plots was not selected according to probabilistic principles, it closely resembled essential properties of the large probability sample. First of all the AGB distributions of the two samples were similar (Figure 6). The mean AGB of the 30 plots with an area of 900 $m^2$ was 366.0 Mg $ha^{-1}$ (Table 4, Figure 6), while it was 461.9 Mg $ha^{-1}$ for the large probability sample (Figure 6). The AGB range was 69.4-908.3 Mg $ha^{-1}$ (standard deviation of 216.3 Mg $ha^{-1}$) while it was 43.2-1147.1 Mg $ha^{-1}$ (standard deviation of 214.7 Mg $ha^{-1}$) for the large sample. Furthermore, the 30 plots covered an elevation range of 200 to 1000 m above sea level (Figure 7a) so that both the lowland forests (<800 m above sea level) and the sub mountain forests (>800 m above sea level) were represented. The 30 plots also covered a wide range of tree sizes (Figure 7b).

### *Field data*

Field data were collected during November 2012, about six mounts after completion of the field work on the large probability sample. On each of the 30 plots, we registered all trees within a radius limited by the maximum distance measuring range of a Vertex hypsometer [36], which was used to measure the horizontal distance from the plot centre to each tree. The maximum measuring range of the hypsometer varied among the plots due to differences in terrain ruggedness and forest density. The radius distribution among the 30 plots was as follows: 31 m (22 plots), 28 m (2 plots), 26 m (1 plot) and 25 m (5 plots). For each tree with diameter at breast height (*dbh*) larger than 5 cm, scientific name, local name, distance to plot centre and *dbh* was registered. A

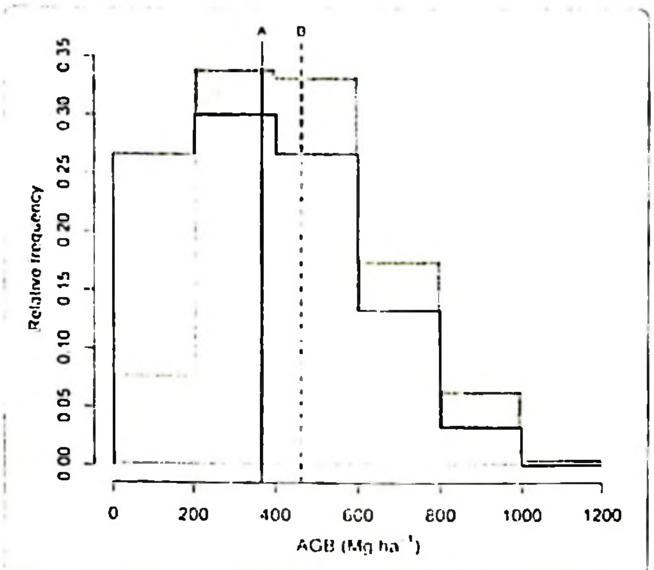



**Figure 6** Distribution of AGB in the large probability sample (light grey), in the small sample of 30 plots [illegible] grey) and overlap between the two distributions [illegible]. The vertical line A indicates the mean of the small sample [illegible] Mg ha⁻¹) and line B the mean of the large sample [illegible] Mg ha⁻¹).

diameter tape, rather than a calliper, was used to gauge diameters since tree trunks in this forest type tend to be both oval and large in size. The distance was measured from plot center to the front of each tree, and half of the tree diameter was added to get the total horizontal distance. The distance measures enabled us to generate any plot size within the limit of the maximum radius. For this study, we decided to select radii between 7.98 m (200 $m^2$) and 30.90 m (3000 $m^2$) (Table 4) for further analysis. Three trees (largest, medium and smallest in terms of diameter) per plot were measured for height ($h$) using a Vertex hypsometer.

Precise field coordinates were determined in the centre of each plot by means of differential Global Navigation Satellite Systems (dGNSS). Topcon Legacy 40 channels dual frequency receivers, observing both pseudo-range and carrier phase of the Global Positioning System (GPS) and the Global Navigation Satellite System (GLONASS) were used as rover and base station. The post-processing reports from Pinnacle version 1.0 software [37] indicated an average error of 19 cm for the planimetric coordinates. The error was computed as two times the standard deviations of the corrected single observations reported from Pinnacle output [38].

### Field estimates of AGB

For each plot AGB was estimated by using the local allometric AGB model developed by [39] with both $dbh$ and $h$ as predictor variables (Eq. 2). Using models with both $dbh$ and $h$ is reported to moderate the effect of large $dbh$-values on AGB estimates as compared to models with $dbh$ only [40-42]. Before calculating AGB, a height model (Eq. 1), was developed using the observations of tree height and corresponding diameters from each plot. A number of model forms for diameter–height relationship [43-48] were tested using non-linear mixed effect approach. Best model fit, judged by the Akaike information criterion (AIC), was obtained using the model form by [46]

$$h = 1.3 + 45.5103 \cdot \left[exp(-2.7163 \cdot exp(-0.0354 \cdot dbh))\right] \quad (1)$$

This model was used to predict height for trees without height measurements. AGB was calculated for individual trees within each plot according to [39] i.e.,

$$AGB = 0.4020 \cdot (dbh)^{1.4365} (h)^{0.8613} \quad (2)$$

and then summed to obtain total AGB for the respective plot. The AGB values were finally scaled to per ha values for the different plot sizes (Table 4). The calculated AGB values are henceforth denoted field reference AGB.

### Laser scanner data

ALS data were collected during the period from 19 January to 18 February 2012 using a Leica ALS70 sensor (Leica Geosystems AG, Switzerland) carried by a Cessna 404 fixed-wing aircraft. Mean flying altitude was 800 m above ground covering the entire area of ANR (i.e. wall to wall) at a ground speed of 75 m $s^{-1}$. The scanning rate was 58.6 Hz and the instrument operated at a pulse repetition frequency of 339 kHz with a resulting average pulse density of 10.6 points $m^{-2}$.

Processing of the ALS data started with classification of each ALS echo as ground or vegetation using the progressive irregular triangular network densification method [49] implemented in the TerraScan software [50]. A Triangular Irregular Network (TIN) was created using the ALS echoes classified as ground echoes. The heights above the ground surface were calculated for all echoes by subtracting the respective TIN heights from the height values of all echoes recorded. Up to five echoes were registered per pulse and we used the three echo categories classified as "single", "first of many", and "last of many". The "single" and "first of many" echoes were pooled into one dataset denoted as "first" echoes, and correspondingly, the "single" and "last of many" echoes were pooled into a dataset denoted as "last" echoes.

Several variables were extracted from the ALS data for each of the field plot sizes as described by [51]. For each plot size, height distributions of both first and last echoes were first created. A height threshold of 2.0 m was applied in order to remove the effect of low vegetation and echoes from ground features falsely classified as vegetation. Then, heights at nine percentiles ($10^{th}$,

**Table 4 Summary of field data**

| Plot size ($m^2$) | Number of plots (n) | Mean (Mg $ha^{-1}$) | Standard deviation (Mg $ha^{-1}$) | Minimum (Mg $ha^{-1}$) | Maximum (Mg $ha^{-1}$) |
|---|---|---|---|---|---|
| 200 | 30 | 411.4 | 323.2 | 55.3 | 1174.5 |
| 300 | 30 | 401.6 | 257.3 | 46.2 | 876.0 |
| 400 | 30 | 424.5 | 275.8 | 72.6 | [illegible] |
| 500 | 30 | 413.8 | 263.4 | 77.6 | 1145.0 |
| 600 | 30 | 395.3 | 243.9 | [illegible] | [illegible] |
| 700 | 30 | 371.5 | 221.5 | 75.4 | 931.7 |
| 800 | 30 | 363.1 | 204.4 | 74.4 | 824.3 |
| 900 | 30 | 366.0 | 216.3 | 69.4 | 903.3 |
| 1000 | 30 | 367.1 | 210.1 | 62.4 | 859.2 |
| 1100 | 30 | 365.6 | 203.0 | 65.4 | 839.5 |
| 1200 | 30 | 365.0 | 193.2 | 76.4 | 797.6 |
| 1300 | 30 | 361.0 | 190.5 | 82.1 | 757.9 |
| 1400 | 30 | 352.3 | 184.2 | 67.3 | 707.0 |
| 1500 | 30 | 354.2 | 180.4 | 55.5 | 757.8 |
| 1600 | 30 | 353.2 | 174.1 | 92.2 | 725.5 |
| 1700 | 30 | 355.0 | 170.2 | 95.6 | 702.6 |
| 1800 | 30 | 355.9 | 163.9 | 91.5 | 696.5 |
| 1900 | 30 | 351.1 | 159.6 | 92.7 | 703.3 |
| 2000 | 25 | 352.2 | 170.8 | 87.6 | 669.3 |
| 2100 | 25 | 350.4 | 168.0 | 85.5 | 646.2 |
| 2200 | 24 | 344.7 | 169.3 | 89.3 | 631.1 |
| 2300 | 24 | 343.0 | 167.8 | 88.5 | 639.8 |
| 2400 | 24 | 344.2 | 171.3 | 97.9 | 677.7 |
| 2500 | 22 | 337.1 | 175.0 | 84.4 | 661.5 |
| 2600 | 22 | 334.1 | 183.1 | 91.8 | 669.9 |
| 2700 | 22 | 326.0 | 179.8 | 88.6 | 674.7 |
| 2800 | 22 | 322.7 | 177.2 | 85.4 | 665.9 |
| 2900 | 22 | 323.5 | 177.9 | 82.5 | 655.6 |
| 3000 | 22 | 321.0 | 179.7 | 79.2 | 666.7 |

Number of plots for the different plot sizes together with mean field reference AGB values with corresponding standard deviation, minimum, and maximum

$20^{th}$, ..., $90^{th}$) of both the first- and last echo distributions were computed to represent canopy height and labeled $H_{10}$.F, $H_{20}$.F, ...., $H_{90}$.F (first echoes) and $H_{10}$.L, $H_{20}$.L, ..., $H_{90}$.L (last echoes), respectively. Measures of canopy density were also derived for first and last echoes of each plot size. The range between the lowest ALS canopy height (>2 m) and the $95^{th}$ percentile height was divided into 10 vertical fractions of equal height. Canopy densities were then computed as the proportion of ALS echoes above each fraction to total number of first echoes and labeled $D_0$.F (>2 m), $D_1$.F, ..., $D_9$.F. Density variables for the last echo distribution were calculated the same way (relative to total number of last echoes) and labeled $D_0$.L, $D_1$.L, ..., $D_9$.L. Furthermore, for both first and last echo height distributions on each plot, the maximum height ($H_{max}$.F and $H_{max}$.L ), mean values ($H_{mean}$.F and $H_{mean}$.L), standard deviation ($H_{sd}$.F and $H_{sd}$.L), coefficient of variation ($H_{cv}$.F and $H_{cv}$.L), and skewness ($H_{skewness}$.F and $H_{skewness}$.L) were computed.

### Data analyses

#### *Model development*

Multiple linear regression analysis with ordinary least square regression (OLS) was used to develop the statistical models relating the field reference AGB and the predictor variables from the ALS data. To ensure that our modelling approaches met the basic assumptions of OLS, the response variable was transformed to logarithmic scale [11,52], while for the predictors both log transformed and non-transformed variables were used. Separate models with log transformed response and combination of log transformed and non-transformed

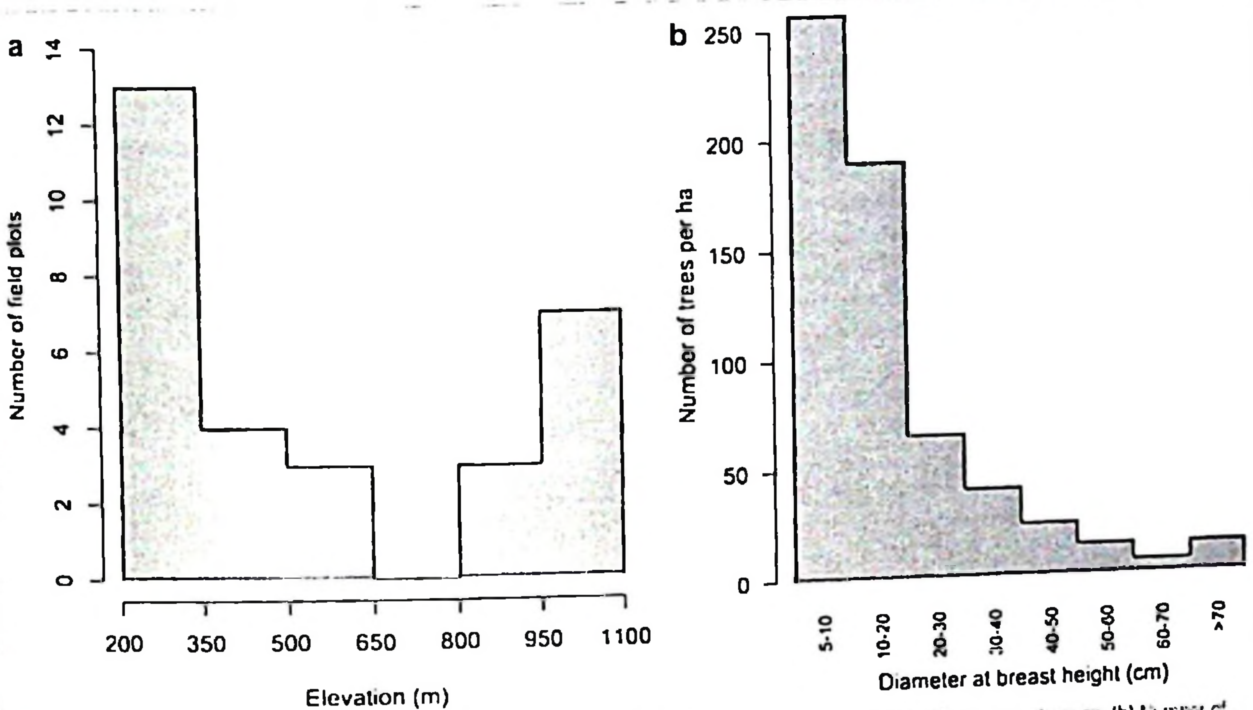



**Figure 7** Distributions of field plots, elevation, number of trees per ha and tree sizes. **(a)** Number of field plots versus elevation. **(b)** Number of trees per ha versus tree sizes.

predictor variables were fitted for each of the plot sizes. We decided to fit separate models (unique variable combinations) for each of the plot sizes, because we wanted the model for each plot size to be the "best" and not be constrained by forcing specific variables into the model.

Variable selection was conducted by using reg-subset in the leaps package in R [53]. The selection of the variables was limited to the best combinations of three or fewer variables in order to avoid multicollinearity among candidate predictors. The preferred models were chosen based on the Bayesian information criterion (BIC) [54]. Adjusted $R^2$ was also used for assessing the model fit while multicollinearity was assessed by computing the variance inflation factors (VIF). The VIF values were determined for the individual $\beta$ parameters. VIF values greater than 10 were regarded as an indication of multicollinearity problems [55].

Log-transformation of the response variable introduces a bias when back-transforming to the arithmetic scale. The model for AGB was therefore adjusted for logarithmic bias according to [56] by adding half of the model mean square error to the constant term before transformation to arithmetic scale.

### *Model validation and accuracy assessment*

In order to assess the performance of the models for each plot size, leave-one-out cross–validation (LOOCV) was performed. One field plot at a time was excluded from the dataset, and the model was fitted based on n-1 plots to predict the AGB of the left out plot. Here, n denotes the number of field plots, where i = 1, ..., n. Relative root mean square error (RMSE %) and the mean prediction error (MPE%) were used as the measures of reliability and calculated according to

$$\mathrm{RMSE\%} = \frac{\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/n}}{\bar{y}} \times 100 \qquad (3)$$

$$\mathrm{MPE\%} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)/n}{\bar{y}} \times 100 \qquad (4)$$

Where $y_i$ and $\hat{y}_i$ denote field reference AGB and predicted AGB for plot i, respectively, and $\bar{y}$ denotes mean field reference AGB for all plots. RMSE% is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction [57].

### *Analysis of boundary effects*

To analyze the boundary effects we studied how the residual errors of the models were related to the field reference AGB of the trees in an outer buffer zone for different field plot sizes. To archive this, we extracted field reference AGB values for 3 m and 6 m buffers outside the field plots for the plot sizes of 200–1500 $m^2$ and

200–1100 $m^2$, respectively. We selected the trees with $dbh > 10$ cm and computed AGB per hectare for the largest tree in the buffer and the total AGB per hectare for all trees in the buffer. To obtain the model residual error, we first subtracted the ground reference AGB from the predicted AGB. Then we calculated the ratio between the residuals and the total field reference AGB for the respective plot (i.e., relative residual). Similar ratios between (1) sum of AGB per hectare for all trees in the buffer ($SAGB_{buffer}$) and the field reference AGB for the plot and (2) the maximum AGB per hectare for the largest tree in the buffer ($MAGB_{buffer}$) and the field reference AGB for the plot were also computed. Two empirical models explaining the variation in the relative residual values using either $SAGB_{buffer}$ or $MAGB_{buffer}$ as explanatory variables were developed. Linear mixed effects (LME) regression using *nlme* add-on package [58] in R was used for model fitting. LME models are linear regression models in which parameters are the sum of the fixed and random effects. In this case the fixed effects were either $SAGB_{buffer}$ or $MAGB_{buffer}$ while plot identity was treated as the random effect. We assumed that each plot will have different random error structures and that the distribution of AGB within these plots is not independent of one another. To test the effect of plot sizes on relative residual, we also fitted the linear regression model which relates relative residuals in absolute form and plot sizes. Absolute value was used because we were interested in the magnitude of the residual regardless of its sign.

### *Efficiency of ALS-assisted AGB estimation*

ALS-assisted estimation of AGB within the design-based and model-assisted inferential framework can greatly improve the precision compared to pure field-based estimation. The purpose of this analysis was to quantify the gain in estimated precision of using ALS data relative to a pure field-based estimate for increasing plot sizes.

A basic requirement for validity of design-based inference is the availability of a probability sample [59]. As stated above, the current sample of 30 plots was obtained as a subsample of a probability sample, but the sub-sampling was not conducted according to strict probabilistic principles. However, the sub-sample was selected to resemble important properties of the large probability sample as closely as practically feasible. Thus, a comparison of variances using the current data and assuming a probabilistic design will most likely introduce a bias in the estimators of unknown magnitude. Likewise, when a systematic sample is obtained, it is common to adopt design-based estimators assuming e.g. simple random sampling (SRS) although it is well-known that SRS variance estimators usually are positively biased under systematic sampling. The magnitude of the bias is always unknown for a particular sample because bias is a property of an estimator and not a particular sample. The current analysis was conducted under the assumption that the sample at hand would give a meaningful quantification of the effect of plot size on relative variance estimates. Thus, in the current study we adopted design-based variance estimators assuming simple random sampling and complete cover of ALS data.

Assuming SRS, the variance estimator for the field-based AGB estimate ignoring corrections for finite population is [60].

$$\hat{V}_{field} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n(n-1)} \tag{5}$$

For model-assisted estimation, the variance estimator of the so-called generalized regression estimator is [60].

$$\hat{V}_{ALS} = \frac{\sum_{i=1}^{n}(\hat{e}_i-\bar{e})^2}{n(n-1)} \tag{6}$$

where $\hat{e}_i = y_i - \hat{y}_i$ is the model prediction residual for plot i and $\bar{e} = \frac{\sum_{i=1}^{n}\hat{e}_i}{n}$ is the mean residual for all plots. Standard error (SE) was computed as the square root of the variance estimates. Finally, the relative efficiency (RE) of ALS-assisted inventory relative to field-based inventory was calculated for different plot sizes as the ratio of the two variance estimates, i.e.,

$$RE = \hat{V}_{field} \Big/ \hat{V}_{ALS} \tag{7}$$

Values of RE greater than 1.0 indicates higher efficiency of ALS-assisted estimates than field-based estimates for a given plot size. To achieve consistency in the analysis across different plot sizes, the dataset was divided into two major groups. The first group subject to analysis comprised all the 30 plots and allowed consistent analysis of plot size ranging from 200–1900 $m^2$. The second group allowing analysis from 200 to 3000 $m^2$ consisted of 22 of the plots.

**Abbreviations**

ALS: Airborne laser scanning; ABA: Area based approach; AGB: Aboveground biomass; AIC: Akaike information criterion; ANR: Amani nature reserve; BIC: Bayesian information criterion; dGNSS: differential Global Navigation Satellite Systems; GLONASS: Global Navigation Satellite System; GPS: Global Positioning System; LME: Linear mixed effects; LOOCV: leave one out cross-validation; MA: Model assisted; $MAGB_{buffer}$: Maximum AGB per hectare for the largest tree in the buffer; MPE: Mean prediction error; MRV: Monitoring, Reporting and Verification; OLS: Ordinary least square regression; REDD+: Reducing Emissions from Deforestation and Forest Degradation in tropical countries; RMSE: Root mean square error; $SAGB_{buffer}$: Sum of AGB per hectare for all trees in the buffer; SAR: Synthetic aperture RADAR; SE: Standard error; SRS: Simple random sampling; TIN: Triangular Irregular Network; UNFCCC: United Nations Framework Convention on Climate Change; RE: Relative efficiency

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

All the authors have made substantial contribution towards successful completion of this manuscript. Authors EWM and OMB have been involved in designing the study, drafting the manuscript, data analysis and write up. EHH has been involved in data analysis and quality control of both raw data and results. EN and TG have been responsible for designing the ALS acquisition and they were involved in revising the manuscript. REM was involved in critical discussion on the field inventory design and all logistics related to the field data acquisition. All authors read and approved the final manuscript.

## Author's information

EWM and EHH are PhD students in forest inventory at Norwegian University of Life Sciences (NMBU). They are both associated with the forest mensuration group in the university. OMB is the researcher in the same group specialized on the application of ALS in forestry. EN and TG are senior scientists and professors in ALS and forest sampling at NMBU. Both EN and TG are resource persons for the forest mensuration group at NMBU. REM is professor in forest inventory and mensuration at Sokoine University of Agriculture, Tanzania.

## Acknowledgements

The financial support for this research was provided by Government of Norway through the two projects entitled "Climate Change Impacts, Adaptation and Mitigation (CCIAM) in Tanzania" and "Enhancing the Measuring, Reporting and Verification (MRV) of forests in Tanzania through the application of advanced remote sensing techniques". We are highly acknowledging our field team in Tanzania and Terratec Norway for collecting and processing of the ALS data. We are also grateful to the administration of AMANI for all support and especially for provision of office space for establishment of the GPS base station.

## Author details

[1]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, Oslo, NO 1432, Ås, Norway. [2]Department of Forest Mensuration and Management, Sokoine University of Agriculture, P.O. Box 3013, Morogoro, Tanzania.

Received: 13 November 2014 Accepted: 29 April 2015
Published online: 07 May 2015

## References

1. Lewis SL, Lopez-Gonzalez G, Sonke B, Affum-Baffoe K, Baker TR, Ojo LO, et al. Increasing carbon storage in intact African tropical forests. Nature. 2009;457:1003–6.
2. Joseph S, Herold M, Sunderlin WD, Verchot LV. REDD+ readiness: early insights on monitoring, reporting and verification systems of project developers. Environ Res Lett. 2013;8:034038.
3. Herold M, Skutsch M. Monitoring, reporting and verification for national REDD+ programmes: two proposals. Environ Res Lett. 2011;6:014002.
4. Keith H, Mackey BG, Lindenmayer DB. Re-evaluation of forest biomass carbon stocks and lessons from the world's most carbon-dense forests. Proc Natl Acad Sci. 2009;106:11635–40.
5. Hyyppä J, Hyyppä H, Leckie D, Gougeon F, Yu X, Maltamo M. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. Int J Remote Sens. 2008;29:1339–66.
6. Vauhkonen J, Maltamo M, McRoberts RE, Næsset E. Introduction to Forestry Applications of Airborne Laser Scanning. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer; 2014. p. 1–16.
7. Coops NC, Wulder MA, Culvenor DS, St-Onge B. Comparison of forest attributes extracted from fine spatial resolution multispectral and lidar data. Can J Remote Sens. 2004;30:855–66.
8. Hansen EH, Gobakken T, Bollandsås OM, Zahabu E, Næsset E. Modeling Aboveground Biomass in Dense Tropical Submontane Rainforest Using Airborne Laser Scanner Data. Remote Sens. 2015;7:788–807.
9. Ioki K, Tsuyuki S, Hirata Y, Phua M-H, Wong WVC, Ling Z-Y, et al. Estimating above-ground biomass of tropical rainforest of different degradation levels in Northern Borneo using airborne LiDAR. For Ecol Manage. 2014;328:335–41.
10. Gautam B, Peuhkurinen J, Kauranne T, Gunia K, Tegel K, Latva-Käyrä P, et al. Estimation of Forest Carbon using LiDAR-Assisted Multi-Source Programme (LAMSP) in Nepal. In: Proceedings of the International Conference on Advanced Geospatial Technologies for Sustainable Environment and Culture. Pokhara, Nepal; 2013. p. 12–3.
11. Næsset E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sens Environ. 2002;80:88–99.
12. Næsset E. Estimating timber volume of forest stands using airborne laser scanner data. Remote Sens Environ. 1997;61:246–53.
13. Næsset E, Bjerknes K-O. Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. Remote Sens Environ. 2001;78:328–40.
14. Næsset E. Area-Based Inventory in Norway: From Innovation to an Operational Reality. In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands: Springer; 2014. p. 215–240.
15. McRoberts RE, Cohen WB, Næsset E, Stehman SV, Tomppo EO. Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scand J For Res. 2010;25:340–67.
16. Frazer GW, Magnussen S, Wulder MA, Niemann KO. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sens Environ. 2011;115:636–49.
17. Gobakken T, Næsset E. Assessing effects of positioning errors and sample plot size on biophysical stand properties derived from airborne laser scanner data. Can J Forest Res. 2009;39:1036–52.
18. Mascaro J, Detto M, Asner GP, Muller-Landau HC. Evaluating uncertainty in mapping forest carbon with airborne LiDAR. Remote Sens Environ. 2011;115:3770–4.
19. Næsset E, Bollandsås OM, Gobakken T, Gregoire TG, Ståhl G. Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data". Remote Sens Environ. 2013;128:299–314.
20. Zolkos S, Goetz S, Dubayah R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sens Environ. 2013;128:289–98.
21. Asner GP, Mascaro J, Muller-Landau HC, Vieilledent G, Vaudry R, Rasamoelina M, et al. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia. 2012;168:1147–60.
22. Gregoire TG, Ståhl G, Næsset E, Gobakken T, Nelson R, Holm S. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. This article is one of a selection of papers from Extending Forest Inventory and Monitoring over Space and Time. Can J Forest Res. 2010;41:83–95.
23. Næsset E, Gobakken T, Solberg S, Gregoire TG, Nelson R, Ståhl G, et al. Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. Remote Sens Environ. 2011;115:3599–614.
24. Ene LT, Næsset E, Gobakken T, Gregoire TG, Ståhl G, Nelson R. Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sens Environ. 2012;123:579–92.
25. Asner GP, Clark JK, Mascaro J, Vaudry R, Chadwick KD, Vieilledent G, et al. Human and environmental controls over aboveground carbon storage in Madagascar. Carbon balance and management. 2012;7:2.
26. Asner GP, Mascaro J. Mapping tropical forest carbon: Calibrating plot estimates to a simple LiDAR metric. Remote Sens Environ. 2014;140:614–24.
27. Mascaro J, Asner GP, Dent DH, DeWalt SJ, Denslow JS. Scale-dependence of aboveground carbon accumulation in secondary forests of Panama: A test of the intermediate peak hypothesis. For Ecol Manage. 2012;276:62–70.
28. Adams T, Brack C, Farrier T, Pont D, Brownlie R. So you want to use LiDAR? a guide on how to use LiDAR in forestry. N Z J For. 2011;55:19–23.
29. White JC, Wulder MA, Varhola A, Vastaranta M, Coops NC, Cook BD, et al. A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. For Chron. 2013;89:722–3.
30. Asner GP. Tropical forest carbon assessment: integrating satellite and airborne mapping approaches. Environ Res Lett. 2009;4:034009.

31. Chen Q, Vaglio Laurin G, Battles JJ, Saah D. Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. Remote Sens Environ. 2012;121:108–17.
32. Gobakken T, Næsset E. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. Can J Forest Res. 2008;38:1095–109.
33. Wulder MA, White JC, Nelson RF, Næsset E, Ørka HO, Coops NC, et al. Lidar sampling for large-area forest characterization: A review. Remote Sens Environ. 2012;121:196–209.
34. Doody K, Howell KM, Fanning E. Amani Nature Reserve. A biodiversity survey. East Usambara Conservation Area Management Programme Technical Paper 52. In: Ministry of Natural Resources and Tourism Tanzania and Frontier Tanzania. Tanga; 2001.
35. Hamilton AC, Bensted-Smith R. Forest conservation in the East Usambara mountains, Tanzania. Gland, Switzerland: IUCN; 1989.
36. Haglöf A. Users guide Vertex III and Transponder T3. Långsele, Sweden: Haglöf Sweden AB; 2002.
37. [illegible] Users Manual. [illegible] Positioning Systems. [illegible] CA, USA; 1999.
38. Næsset E. Effects of differential single- and dual-frequency GPS and GLONASS observations on point accuracy under forest canopies. Photogramm Eng Remote Sens. 2001;67:1021–6.
39. Masota A. Tree allometric models for predicting above- and belowground biomass of tropical rainforests in Tanzania. In press.
40. Feldpausch TR, Banin L, Phillips OL, Baker TR, Lewis SL, Quesada CA, et al. Height-diameter allometry of tropical forest trees. Biogeosciences. 2011;8:1081–106.
41. Banin L, Feldpausch TR, Phillips OL, Baker TR, Lloyd J, Affum-Baffoe K, et al. What controls tropical forest architecture? Testing environmental, structural and floristic drivers. Glob Ecol Biogeogr. 2012;21:1179–90.
42. Mugasha WA, Bollandsås OM, Eid T. Relationships between diameter and height of trees in natural tropical forest in Tanzania. Southern Forests: J For Sci. 2013;75:221–37.
43. Nilsson U, Agestam E, Ekö P-M, Elfving B, Fahlvik N, Johansson U, et al. Thinning of Scots pine and Norway spruce monocultures in Sweden. [illegible]
44. Ratkowsky DA, Giles DE. Handbook of nonlinear regression models. New York: Marcel Dekker; 1990.
45. Richards F. A flexible growth function for empirical use. J Exp Bot. 1959;10:290–301.
46. Winsor CP. The Gompertz curve as a growth curve. Proc Natl Acad Sci U S A. 1932;18:1.
47. Wykoff WR, Crookston NL, Stage AR. User's guide to the stand prognosis model. In: US Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station; 1982.
48. Yang RC, Kozak A, Smith JHG. The potential of Weibull-type functions as flexible growth curves. Can J Forest Res. 1978;8:424–31.
49. Axelsson P. Processing of laser scanner data—algorithms and applications. ISPRS J Photogramm Remote Sens. 1999;54:138–47.
50. Axelsson P. DEM generation from laser scanner data using adaptive TIN models. Int Arch Photo Remote Sensing. 2000;33:111–8.
51. Næsset E. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scand J For Res. 2004;19:164–79.
52. Hudak AT, Crookston NL, Evans JS, Falkowski MJ, Smith AMS, Gessler PE, et al. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. Can J Remote Sens. 2006;32:126–38.
53. Team RC. R: a language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna, Austria. In: ISBN 3-900051-07-0; 2013.
54. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
55. Fox J, Weisberg S. An R companion to applied regression. United Kingdom: Sage; 2011.
56. Goldberger AS. The interpretation and estimation of Cobb-Douglas functions. Econometrica: J Econ Soc. 1968:464–472.
57. [illegible]
58. [illegible]
59. McRoberts RE, Næsset E, Gobakken T. Inference for lidar-assisted estimation of forest growing stock volume. Remote Sens Environ. 2013;128:268–75.
60. Draper NR, Smith H. Applied regression analysis. New York: Wiley; 1998.

# PAPER IV

# Modelling and predicting measures of tree species diversity using airborne laser scanning data in miombo woodlands of Tanzania.

Ernest William Mauya, Ole Martin Bollandsås, Tron Eid, Terje Gobakken and Erik Næsset.

Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

**Abstract**

The decline of biodiversity in tropical forests is one of the major current global concerns. However, there is still lack of information on forest biodiversity for supporting different conservation efforts that aim at protecting and restoring the biodiversity of tropical forests. Current remote sensing techniques may close this gap by providing spatially continuous information that can be used for biodiversity assessment and monitoring. In this study, we evaluated the potential of using remotely sensed data derived from airborne laser scanning (ALS) for modeling and predicting measures of tree species diversity. Specifically, we considered tree species richness and Shannon diversity index since they are commonly used measures of tree species diversity. Two modelling approaches were tested: linear mixed effects modelling (LMM), by which each of the measures was modelled separately, and the *k*-nearest neighbor technique (*k*-NN), by which both measures were jointly modelled (multivariate approach). For both methods, we tested the effect of vegetation type on the prediction accuracies of tree species richness and Shannon diversity index. Separate predictions for richness and Shannon diversity index using LMM resulted in relative root mean square errors ($RMSE_{cv}$) of 40.7%, and 39.1%, while for the *k*-NN they were 41.4% and 39.1%, respectively. Inclusion of dummy variables representing vegetation types to the LMM improved the prediction accuracies of tree species richness (RMSEcv = 40.2%) and Shannon diversity index (RMSEcv = 38.0%). The study showed that ALS data has a large potential for modelling and predicting measures of tree species diversity in the miombo woodlands of Tanzania.

Key words: *Airborne laser scanning, biodiversity, tree species diversity, k-NN.*

## 1.0. Introduction

Tropical forests ecosystems host at least two thirds of the earth's terrestrial biodiversity and provide significant human benefits at local and regional scales through the provision of economic goods and services (Gardner et al., 2009). Nevertheless, the biodiversity of tropical forests is declining rapidly due to conversion of forests to permanent agricultural land, climate change, induced fires, and unsustainable logging practices (Bellard et al., 2012; Chidumayo, 2013). Other consequences of these changes include increasing amount of carbon emission from tropical forests as compared to other forest types globally. To mitigate this development, and to conserve the beneficial ecosystem services that tropical forests can provide, Reducing Emissions from Deforestation and Forest Degradation (REDD+) has been negotiated as an international effort under the United Nations Framework Convention on Climate Change (UNFCCC) since 2005. REDD+ has now become the most active global land use policy program (Matthews and van Noordwijk, 2014) in the tropical countries. Although the primary objective of the REDD+ framework is climate change mitigation through enhancement of forest carbon stock and sustainable forest management, co-benefits from biodiversity conservation are also expected (Imai et al., 2014). However, despite this considerable potential, methods for monitoring forest biodiversity are lacking in the context of REDD+ (Ehara et al., 2014; Gardner et al., 2012; Roe et al., 2013). Efforts have so far focused on the establishment of methods for accounting large-scale forest carbon stock and changes, with little emphasis on forest biodiversity assessment (Imai et al., 2014). Such methods are also important for other subsidiary bodies like the Convention on Biological Diversity (Chandra and Idrisova, 2011) and Millennium Ecosystem Assessment (Mooney et al., 2004).

Assessments of biodiversity for management, conservation planning or policy development (e.g. REDD+), however, poses a number of challenges, especially at larger scales due to the broad, multi-dimensional and multi-scale characteristics associated with such assessments (Boutin et al., 2009). Unlike in forest carbon assessments, where agreed and consistent metrics exist (e.g. carbon density or emissions in metric tons per hectare), there is lack of consensus on what to monitor in biodiversity assessments (Pereira et al., 2013). To account for this, the emphasis has turned to acquiring data for estimation of indicators that aggregate and synthesize information for describing multiple aspects of biodiversity simultaneously (Winter et al., 2011). Species diversity is one of the most intuitive and widely adopted indicators of biodiversity because it is strongly correlated with other biodiversity attributes such as genetic diversity and ecosystem functioning (Chiarucci and Palmer, 2006; Colwell and Coddington, 1994; Pereira and Cooper, 2006). Thus in many studies that aimed at describing biodiversity of forest ecosystems, tree species diversity has been used as an indicator of forest biodiversity (e.g. Barna and Bosela, 2015; Schmidt et al., 2015; Shirima et al., 2015; Vanderhaegen et al., 2015). Additionally, from cost-benefit and time-efficiency perspectives, tree species diversity is considered as a more viable option for biodiversity assessment since it can easily be derived from existing datasets such as national forest inventories (Chirici et al., 2012).

Species richness and species evenness are two basic measures describing species diversity of different taxa (Magurran, 1988). Species richness, computed as the total number of species in a community, has frequently been used as a measure of tree species diversity in different forest types (e.g. Chiarucci and Palmer, 2006; McRoberts and Meneguzzo, 2005), especially for those cases that involves comparisons of conservation values of different sites . Species evenness (sometimes known as equitability) is describing the way abundance is distributed among individual species in a community. A biological community, in which all the species are represented by the same number of individuals, has high species evenness while a community in which a few species are represented by many individuals and the other species are represented by few individuals, has a low species evenness. A number of indices taking into account both species richness and evenness have also been proposed for measuring species diversity. In this category, the Shannon diversity index (e.g. Magurran, 1988) has frequently been used in many studies on tree species diversity (Borah et al., 2015; Nadeau and Sullivan, 2015; Shirima et al., 2015).

The computation and reliability of measures for tree species richness and evenness, or for the two combined, in estimating tree species diversity rely heavily on accurate field-based information. Traditionally assessment of tree species diversity has been done by using ground-based surveys. However, for larger spatial scales ground field based surveys are impractical due to huge amount of data to be collected. Current remote sensing techniques may close this gap by providing spatially continuous and time-series information that can be used to describe tree species diversity (Fricker et al., 2015; Müller and Vierling, 2014). Moreover, remote sensing allows frequently repeated recording of environmental information, and may thereby provide time-series information that are essential for monitoring and understanding how tree species diversity respond to different environmental factors over time (Leutner et al., 2012).

Airborne laser scanning (ALS) is a remote sensing technique that recently has gained wide acceptance in ecologically based studies due to its ability to quantify the three dimensional (3D) structure of forests, which is of particular interest in characterizing measures of tree species diversity and other taxa in the forests ecosystem (Fricker et al., 2015; Müller and Vierling, 2014). However, ALS cannot measure directly the measures tree species diversity, thus applications of ALS involve the development of statistical models or classifiers that relate the ALS structural metrics to measures of tree species diversity derived from field plots. By using such models or classifiers along with statistical sampling estimators (e.g. Chao and Shen, 2003), estimates of tree species diversity may be produced for relevant geographical areas of interest. Models may also be used for developing tree species distribution maps that can support decision-making and conservation planning. Several studies in tempearate and boreal forests, have used ALS data for modelling and predicting measures of tree species diversity (Ceballos et al., 2015; Leutner et al., 2012; Simonson et al., 2012) together with other taxa such as beetles (Müller and Brandl, 2009), spiders (Vierling et al., 2011) and birds (Lindberg et al., 2015; Vogeler et al., 2014). Parametric, and to lesser extent non-parametric methods (e.g. random forests), have been used in these studies.

The non-parametric approach *k*-nearest neighbors (*k*-NN) have widely been used in modelling and estimation of different forest attributes such as volume and aboveground biomass (AGB) when using remotely sensed data (McRoberts et al., 2015). Among the desirable features of *k*-NN are the ability to handle multivariate data and non-linear and diverse relationships between dependent and independent variables (Eskelson et al., 2009). The multivariate feature of the *k*-NN technique makes it very useful in ecological applications because management decisions frequently require consistent information on multiple parameters and estimates. However, to our knowledge no study has to date attempted to use *k*-NN for modelling and prediction of measures of tree species diversity using ALS data. Of particular interests is the tropical forests, where irrespective of the methods used, only a few studies have quantified the relationship between measures of tree species diversity and ALS data (Müller and Vierling, 2014). Lack of ALS data and a complex structure due to the large number of tree species in tropical forests, as compared to temperate and boreal forests, are among the possible reasons for that fewer studies of this kind have been carried out in tropical forests. Furthermore, the few studies that do exist (e.g. Fricker et al., 2015; Hernández-Stefanoni et al., 2014), have not attempted to analyze the influence of vegetation types/forest types on the relationship between measures of tree species diversity and ALS data. Vegetation types are considered as important sources of forest structural variation (Swatantran et al., 2011), which would also affect the prediction accuracy of tree species diversity when using

ALS data. Stratification and post-stratification of forest inventory information has been suggested as a viable means to reduce prediction errors due to structural variations (Latifi et al., 2015), particularly in the studies related to prediction of different forest attributes using ALS data. However, such studies are still limited in the aspects of tree species diversity as compared to AGB and volume.

The overall objective of this study was to assess if ALS data can be used to predict measures of tree species diversity in miombo woodlands, forests and other vegetation types surrounding the woodlands in Liwale district, Tanzania. The specific objectives were to: 1) examine the performance of parametric and non-parametric methods for predicting measures of tree species diversity using ALS data; and 2) assess the prediction accuracy of measures of tree species diversity across vegetation types.

## 2.0. Materials and Methods

### *2.1. Study area description.*

The study site is located in Liwale district Lindi region (Fig. 1) and it occupies an estimated area of about 15.867 km$^2$ in the south-astern part of Tanzania. The altitude ranges from 360 to 900 meters above sea level. The area experiences an annual rainfall ranging from 600 mm to 1000 mm. the majority between November and early April. The average annual temperature ranges between 20$^0$C and 30$^0$C (LDC. 2014). The vegetation type that dominates the study area is miombo woodlands. characterized by the presence of three genera: *Brachystegia*, *Julbernardia*, and *Isoberlinia* from the family *Fabaceae* and sub-family *Caesalpinioideae*. Even though the area is mainly dominated by miombo woodlands. forests and other vegetation types that were neither woodlands nor forests. were also found in our study area.

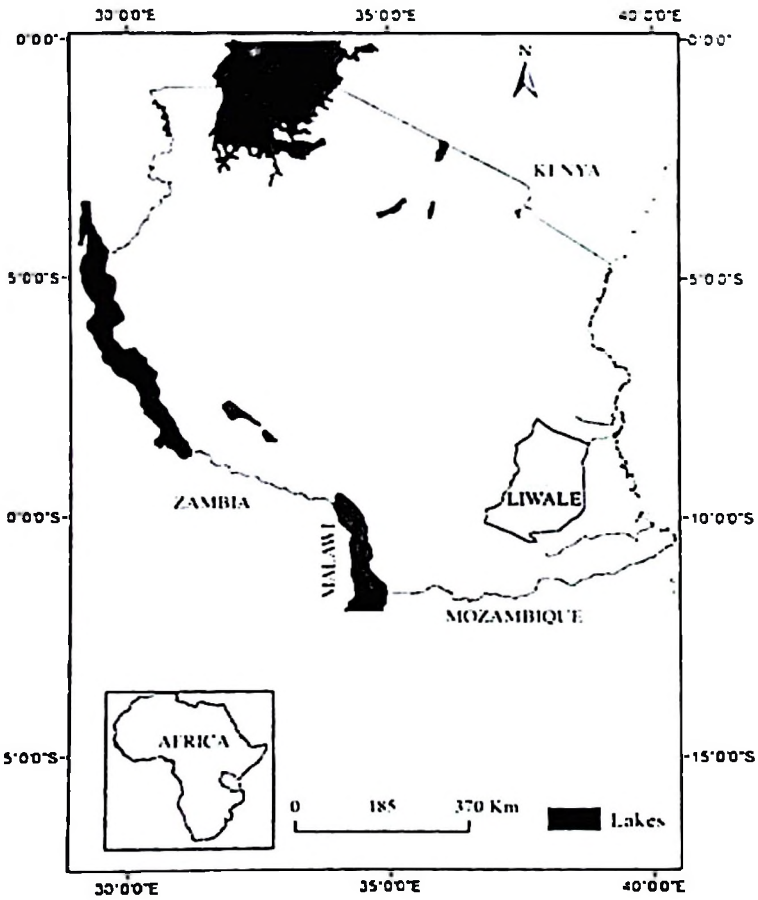



Fig. 1. Location of the study area in Liwale district, southern Tanzania.

### 2.2. *Data collection and processing*

#### 2.2.1. *Field data*

In this study we used the field plots that were established and measured by the National Forest Resources Monitoring and Assessment (NAFORMA) program in 2011. NAFORMA is the first ground based national forest inventory of Tanzania that has been conducted from 2009 to 2014 covering different vegetation types across the country (MNRT, 2015). In our study area (i.e. Liwale) the initial measurement by NAFORMA was completed in June 2011. Eight months later (February 2012) we revisited all the plots to ensure temporal consistency with the ALS data acquisition (February/March 2012) but also to accurately record the positions of the plots using survey grade Global Positioning System and the Global Navigation Satellite System receivers.

#### 2.2.2. *Sampling design*

The sampling design adopted by NAFORMA is systematic double sampling for stratification with individual plots allocated in clusters (Fig. 2). The design was chosen after sampling simulations to reduce uncertainty of estimates under given budget constraints (Tomppo et al., 2014). In the first phase, a dense grid of clusters was over laid on the map of Tanzania mainland at distance of 5×5 km between the clusters. All first phase clusters across the country were then stratified into 18 strata based on three criteria; predicted growing stock volume, accessibility and slope. The second phase sample was systematically selected from the first phase for each individual stratum with different sampling intensities in each of the 18 strata. Higher sampling intensities were used in strata with large predicted growing stock volume and lower sampling intensities were applied in strata with low predicted growing stock. The clusters selected during the second sampling phase determined which clusters to be measured in the field. Ten plots were allocated systematically according to an "L"-shaped pattern in each cluster (Fig. 2). The distance between plots within a cluster was 250 m while the distance between clusters varied from 5 to 45 km. The details of the planning of this design are given in Tomppo et al. (2014).

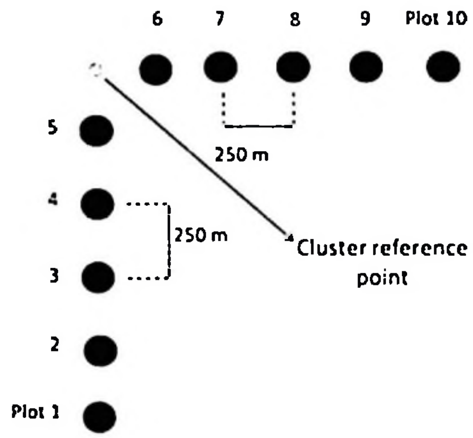


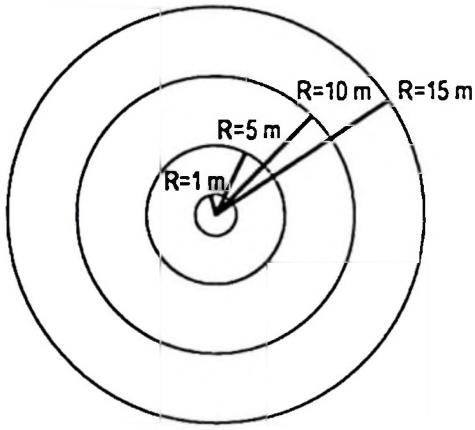

Fig. 2. The structure of a NAFORMA cluster. The field sample plots are represented by black dots, and the ALS coverage is shown with gray shading. Note that the ALS measurements do not cover the entire cluster.

### *2.2.3. Plot data*

The NAFORMA cluster structure has a total of ten plots per cluster (MNRT, 2011), however during the re-measurement we omitted two plots in each cluster (Fig. 2), because they were outside the corridors designated for ALS data acquisition. Thus, eight plots per cluster were re-measured in the field. Differential Global Navigation Satellite Systems (dGNSS) were used to positions the center of each sample plot. Two Topcon Legacy 40-channels dual frequency receivers observing both pseudo range and carrier phase of the Global Positioning System (GPS) and the Global Navigation Satellite System (GLONASS) were used as rover and base station, respectively.

The re-measurements of the NAFORMA plots used in the current study followed the NAFORMA field protocol. The plots had a concentric circular design with 15 m radius. Within each plot, diameter at breast (*dbh*) for the trees was measured using calipers or diameter tape following the concentric plot design described in Fig. 3. Height measurements were done for every fifth tally tree in the cluster by using a Suunto hypsometer. Identification of species names of every recorded tree in the plot was done by the professional botanist from Tanzania Forest Research Institute (TAFORI). The identification was also guided by the previous tree species information recorded by NAFORMA in 2011 to ensure consistency between the two measurements and that the same names were used for every measured tree. For all the trees that were identified, the species names were further confirmed by using NAFORMA checklist to ensure that botanic names are correctly recorded, this was done in the Herbarium at Lushoto Silviculture Research Centre before further analyses.

On each plot, information on vegetation types were also recorded. For this study, we grouped our data into three distinct vegetation types (NAFORMA broad categories): woodlands, forest, and other vegetation types as described in Table 1 and further elaborated in MNRT (2011). This information was used to stratify the plots into more homogeneous ecological groups with the objective of helping to explain the variation in tree species diversity better.




1) Within 1 m radius; all trees with dbh > 1cm were recorded

2) Within 5m radius; all trees with dbh > 5cm were recorded

3) Within 10m radius; all trees with dbh >10 were recorded

4) Within 15 m radius; all trees with dbh > 20cm were recorded

Fig. 3. NAFORMA concentric sample plot design.

Table 1. Brief description of the structure and biological values of the three vegetation types.

| Vegetation types | Description | Biodiversity |
|---|---|---|
| Forest | The forest category referred to in this study comprised of low land forests.<br>• They occur at an elevation less than 800m.<br>• They are characterize by semi ever green closed and dense canopy. | High biodiversity values. |
| Woodlands | • Generally, the canopy cover for the woodlands is less dense as compared to forest.<br>• Trees are well spaced with short trunked and spreading canopies.<br>• Most of the woodlands are deciduous. | Moderate biodiversity value. |
| Other vegetation types | • In this category, we refer to bushland, grassland and cultivated land.<br>• The canopy cover in these categories is more sparse as compared to forest and woodlands.<br>• Detail description of individual categories are given in MNRT (2011). | Less biodiversity value. |

*2.2.4 Calculation of the measures of tree species diversity*

In this study we considered tree species richness and Shannon diversity index as the commonly used measures of tree species diversity. Calculation of the two measures of tree species diversity was done for each sample plot, and only trees with dbh ≥ 5 were considered. Tree species richness (S) was calculated by summing the number of tree species in each of the plot $i$, while Shannon diversity index ($H'$) was calculated as:

$$H' = -\sum_{i=1}^{S} p_i \ln p_i \qquad (1)$$

where $pi$ is the proportion abundance of the $i$th species relative to the total abundance of all the species ($S$) sampled in the plot $i$ and $ln$ is the natural logarithm. Shannon diversity index assumes that the $p_i$'s are population parameters where all the species in the population are known. However, in practice, the reported Shannon diversity index is only an estimator of the population and is biased because the number of species observed in the sample (e.g. field plot) is less than species in the population. But, if sampling is adequate, this bias is considered to be small (Oldeland et al., 2010).

To assess sampling adequacy of our sample plots, we developed species accumulation curves (Kindt and Coe, 2005) as presented in Appendix 1 and 2. The curves seem to approach asymptotic levels, indicating that the number of sample plots employed provided adequate representation of tree species richness of the study area. All the calculations and plotting were done by using vegan library (Oksanen et al., 2011) in R statistical software following the procedure by Kindt and Coe (2005). Summary statistics for the tree species richness, and Shannon diversity index over the different vegetation types are presented in Table 2.

Table 2. Number of field sample plots (n), total tree species richness, mean, standard error (Se) of tree species richness and Shannon diversity index for different vegetation types.

| Vegetation type [a] | n | Tree species richness | | | Shannon diversity index | |
|---|---|---|---|---|---|---|
| | | Total | Mean | Se | Mean | Se |
| Woodlands | 387 | 239 | 5.36 | 0.12 | 1.20 | 0.02 |
| Forest | 40 | 90 | 6.68 | 0.36 | 1.29 | 0.07 |
| Agriculture and other cover types | 57 | 79 | 3.23 | 0.31 | 0.73 | 0.08 |
| All | 484 | 270 | 5.22 | 0.12 | 1.15 | 0.02 |

[a] Vegetation types according to MNRT (2011).

### 2.2.5. *ALS data*

The ALS data were acquired in the period between 10 February and 7 March 2012. Thirty-two parallel strips with an average width of 1374 m, were systematically distributed over the study area in east-west direction (Fig. 4). The ALS strips were spaced 5 km apart, following the NAFORMA 5×5 km grid. A Leica ALS 70 airborne laser sensor (Leica Geosystems AG, Switzerland), carried by a Cessna 404 aircraft was used for data acquisition. The measurements were acquired from an average flying altitude of approximately 1200 m above the ground, at an average ground speed of 77.2 $ms^{-1}$. The sensor was operated at a pulse repetition frequency of 193 kHz, with a scan rate of 36.5 Hz. The beam divergence was 0.28 m rad which produced an average footprint size on the ground of about 34 cm. The average pulse density was 1.8 $m^{-2}$.

The data were initially processed by the contractor (TerraTec AS, Norway), where the first step was to classify echoes into ground- and vegetation echoes. Then, a terrain model was built using the progressive Triangular Irregular Network (TIN) densification algorithm (Axelsson, 2000) implemented in TerraScan software (Anon 2012). The heights above ground (the TIN surface) were calculated for all vegetation echoes by subtracting the respective xy-corresponding TIN heights from the echo height values. Up to four echoes were registered per pulse and we used the three echo categories classified as "single", "first of many", and "last of many". The "single" and "first of many" echoes were pooled into one dataset denoted as "first" echoes, and correspondingly, the "single" and "last of many" echoes were pooled into a dataset denoted as "last" echoes.

Separate height distributions were derived from the first and last ALS echoes and used to extract ALS metrics for each sample plot. A threshold of 1.3 m above ground was used to separate canopy echoes. Below this height, echoes were considered to have been reflected from shrubs, grass, or ground, i.e., non-tree objects. Height variables for first and last echoes including maximum value (MaxF and MaxL), mean value (MeanF and MeanL), coefficient of variation (CVF and CVL), and percentiles at 10% intervals labeled PF0, PF10, ..., PF90 (first echoes) and PL0, PL10, ..., PL90 (last echoes), were computed. Furthermore, several measures of canopy density were derived. Canopy density was calculated for 10 different vertical layers according to Næsset (2004). The height of each layer was defined as one tenth of the distance between the 95 percentile and the lowest canopy height, i.e., 1.3 m (Gobakken et al., 2012). Canopy densities were then computed as the proportions of ALS echoes above fraction #0, 1, ..., 9 to total number of echoes, and denoted TF0 ( >1.3 m), TF1, ..., TF9 for the first echoes and TL0, TL1, ..., TL9 for the last echoes.

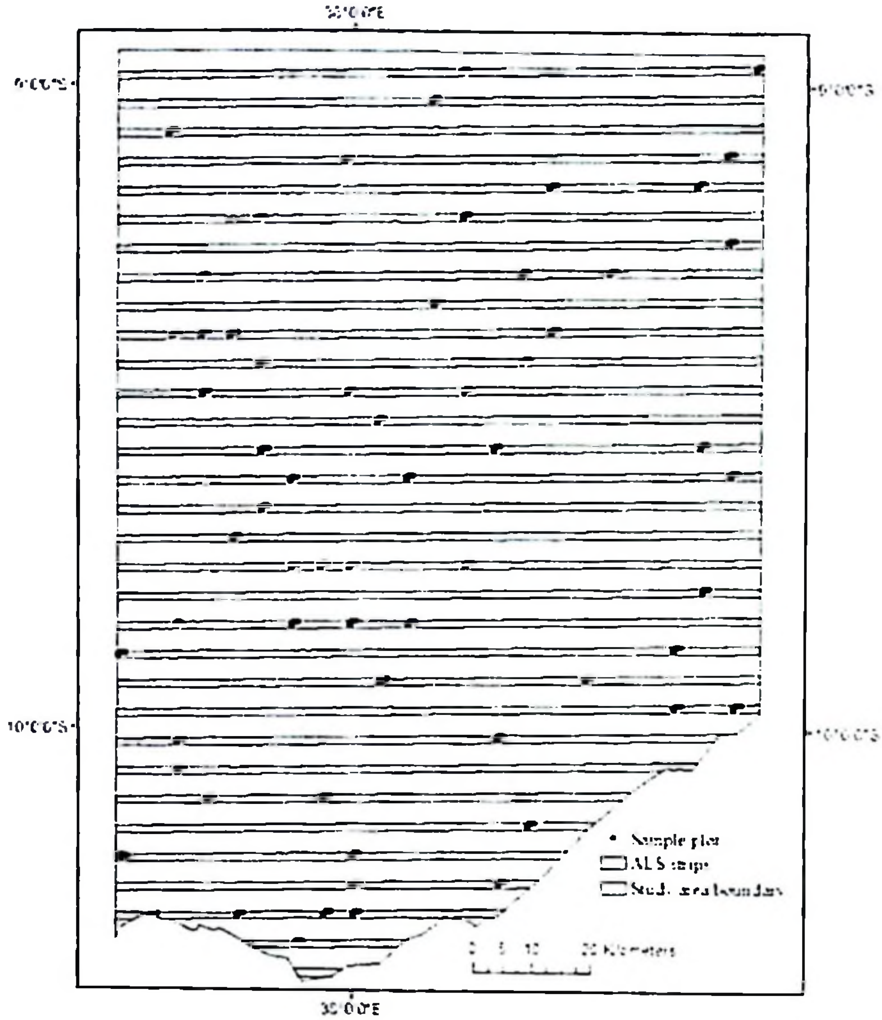


Fig. 4. Map of the study area, ALS strips coverage, and distribution of sample plots.

## 2.3. *Statistical analyses*

### 2.3.1. *Outline of the analyses*

Parametric and non-parametric methods were used to develop statistical relationship between the measures of tree species diversity and the ALS metrics. Specifically we used linear mixed effects models (LMMs) and the *k*-NN approach. The analyses were performed based on the following steps below:

1. Explanatory variables for predicting each of the measures of tree species diversity (i.e. tree species richness and Shannon diversity index) were selected separately for fitting LMMs.
2. Separate LMMs for tree species richness and Shannon diversity index were fitted.
3. Multivariate approach was then used for selecting explanatory variables and predicting measures of tree species diversity using *k*-NN techniques.
4. Both of the two methods were evaluated and compared using measures of reliability derived from the cross validation.
5. Effects of vegetation types on the prediction accuracy of the measures of tree species diversity for each of the method was assessed.

### *2.3.2. Parametric method*

*Model development (LMMs)*

When modelling measures of tree species diversity using ALS data previous studies (e.g. Fricker et al., 2015; Hernández-Stefanoni et al., 2014; Wolf et al., 2012) ordinary least square regression (OLS) have most frequently been used. However, in this study the sampling design characteristics employed by NAFORMA (section 2.2.2) imposes a hierarchical data structure where the field plots are nested within the clusters. Theoretically this may induce a spatial dependency among the plots measured from the same cluster, and thus it is likely that the basic assumptions of uncorrelated error terms might not hold. LMM in this case was potentially an ideal tool for development of predictive models that account for dependence of the plots within the clusters, but also for ensuring that the modelling procedure adheres to the sampling design. LMM essentially consists of two main parts i.e. a fixed effect- and random effect part. The fixed effects are common to all subjects, while random effect parameter are specific to each subject (Pinheiro and Bates, 2000) .

Model development for each of the measures of tree species diversity started with variable selection, i.e. separate explanatory variables were selected for richness and Shannon diversity index. In both cases a subset regression technique was used for variable selection. Specifically we used *reg-subsets* function implemented in the leaps Package of the R software (Team, 2014). The model statistics used to determine the best subsets was Bayesian Information Criterion (BIC). The selection of the variables was limited to the best combinations of four or fewer variables in order to avoid multicollinearity among candidate predictors. We also computed VIF values for each of the parameters ($\beta$s), and values greater than 10 were considered as an indication of multicollinearity. Variable combinations that yielded VIF-values higher that this threshold were excluded from the model (O'brien, 2007).

To account for heteroscedasticity and non-linearity, the model for tree species richness was fitted with natural-log transformation of the response variable and non transformed predictor variables. Such model form has been used by Hernández-Stefanoni et al. (2014) with good results when modelling and predicting tree species richness using ALS data. For Shannon diversity index we used non transformed model. All the models were fitted using restricted maximum likelihood estimation (REML) procedure in lme4 package (Pinheiro et al. 2007) of the R software (Team 2014). Quality of the model fits (coefficient of determination) were assessed by using pseudo R-square ($R^2$) computed as the square of the Pearson correlation coefficient between observed and predicted values. The standard form of LMMs as applied for each measure of tree species diversity are presented below:

$$\ln S_{ij} = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_k x_{ijk} + u_i + \epsilon_{ij} \quad (2)$$

$$H'_{ij} = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_k x_{ijk} + u_i + \epsilon_{ij} \quad (3)$$

$$i = 1, \ldots\ldots, n_i \qquad j = 1, \ldots\ldots, m \quad u_i \sim N(0, \sigma_\mu^2) \quad \epsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where $S_{ij}$ and $H'_{ij}$ are tree species richness and Shannon diversity index of $i$th sample plot in the $j$th cluster, $x_{ij1}, \dots\dots\dots, x_{ijk}$ are $k$ fixed effects, $\beta_0, \dots\dots\dots, \beta_k$ are the fixed effects parameters, $n_i$ is the number of sample plots within the cluster $j$, $m$ is the number of clusters. We assumed that cluster level random effects $\mu_i$ were independent of the plot level residuals $\varepsilon_{ij}$.

*Accuracy assessment*

In order to assess the prediction accuracy of tree species richness and Shannon diversity index when using ALS data, we applied leave one out cross validation (LOOCV) at the cluster level to ensure that hierarchical data structure was preserved during re-sampling (i.e. leave one cluster out). The predicted values from LOOCV were corrected for biasness due to logarithmic transformation using the method suggested by Snowdon (1991). The accuracies of the prediction from the LOOCV for tree species richness and Shannon diversity index were evaluated by using relative root mean square error (RMSEcv%):

$$\text{RMSE}_{\text{CV}}\% = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2/n}}{\bar{y}} \times 100 \tag{4}$$

where $y_i$ and $\hat{y}_i$ denote observed and predicted tree species richness and Shannon diversity index respectively, for plot $i$, and $\bar{y}$ denotes their mean field observed value for all plots.

*2.3.3. Non-parametric method*

*k*-NN *imputation*

Non-parametric methods have gained wide acceptance in ecologically based studies given their unique ability to account for complex relationships and spatial patterns as compared to the traditional probability based approaches (Drew et al., 2010). Furthermore, these methods allows univariate and multivariate predictions of both continuous and categorical variables. In this study, we used the *k*-NN imputations method carried out with the package *yaImpute* (Crookston and Finley, 2008). In *k*-NN language and set up the dataset should be distinguished between reference and target sets. The population units for which observations of both response and explanatory variables are available is labeled reference set; the set of the population units for which only the explanatory variables are available is termed as the target set. In our study, the reference set contained both measures of tree species diversity (i.e. tree species richness and Shannon diversity index) and the ALS metrics, while the target set contained only the ALS metrics.

In typical *k*-NN imputation, the dependent variable for the target observation is predicted by means of finding its *k* nearest neighbor observations in the reference dataset and assigning the value of the variable to be the weighted averages of the of the values of the neighbors. Nearness of the observations is measured with the independent variables and is defined in terms of weighted Euclidean distance. The principle behind *k*-NN imputation as it is applied in this study is further explained in Eskelson et al. (2009); McRoberts (2012) and its use for forest parameter estimation in *yaImpute* statistical environment can be found in Hudak et al. (2008).

In order to identify predictor variables for predicting tree species richness and Shannon diversity index simultaneously, multivariate variable selection was firstly done by using the *VarSelection* function in the R *yaImpute* package. Model fitting and imputation were then performed by using *yai* and *impute tools* within the *yaImpute* package. In this package, any *k* number of reference observations can be selected to impute the target. We tested the values of *k* from 1 to 10 and selected the value with lowest RMSEcv from cross validation. We used LOOCV at the cluster level, where one cluster at time was used as the target set while the remaining clusters were used at the reference set. The imputed values from the LOOCV were used to compute RMSEcv% as described in equations 4 and used to compare with LMM.

### *2.3.4. Effect of vegetation types on the prediction accuracy*

To account for the variability in prediction accuracy that might be caused by the differences in vegetation types, we firstly fitted and evaluated the LMMs that relates each of the measures of tree species diversity and ALS data, with dummy variable that specify vegetation types. Secondly vegetation specific LMMs for each of the measures of tree species diversity were fitted and evaluated. We applied the same procedure described above for variable selection and accuracy assessment. Lastly multivariate *k*-NN imputation with $k = 1$ was also applied for each of the vegetation type. RMSEcv% from the LOOCV for each of the method was calculated to assess the variability in prediction accuracy across different vegetation types.

## 3.0. Results

### *3.1. Parametric method*

LMMs for predicting tree species richness and Shannon diversity index were tested and evaluated. Predictor variables derived from both the first and last echoes were selected from the best subset regression. The number of predictors for the individual models were four consisting of both canopy height and canopy density metrics. For all the selected variables, the VIF values were < 10, indicating acceptable levels of multicollinearity. Similarly the parameter estimates for LMMs fitted for each of the measures of tree species diversity (Table 3), were significantly different from zero ($p < 0.05$). The LMMs explained relatively more of the variation in tree species richness as compared to Shannon diversity index. However, results from the LOOCV indicated lower RMSEcv% values for Shannon diversity index as compared to tree species richness (Table 4). A more detailed evaluation of the prediction accuracy of the measures of tree species richness and Shannon diversity index is presented using the validation plots (Fig. 5), which shows the relationship between predicted and corresponding observed values.

Table 3. Parameter estimates of LMMs for tree species richness and Shannon diversity index.

| Tree species richness | | | Shannon diversity index | | |
|---|---|---|---|---|---|
| Predictor variables | Parameter estimates | Standard error | Predictor variables | Parameter estimates | Standard error |
| Intercept | 0.6372 | 0.0872 | Intercept | 0.4850 | 0.0788 |
| MaxF | 0.0346 | 0.0072 | MaxF | 0.0262 | 0.0064 |
| PF20 | -0.0782 | 0.0115 | PF10 | -0.0518 | 0.0123 |
| TL0 | -0.8061 | 0.1769 | TF2 | 1.4667 | 0.1894 |
| TF2 | 2.0465 | 0.2002 | TL0 | -0.7453 | 0.1662 |

PF20 = Percentiles of the first echo canopy heights for 20% (m); TF2 = Canopy densities corresponding to the proportion of first echoes above fraction #2; TL0 = Canopy densities corresponding to the proportion of last echoes above fraction # 0 (1.3m), MaxF = Maximum of the canopy height distributions of the first echoes

Table 4. Pseudo-$R^2$ ($R^2$), absolute root mean square error (RMSE), and relative root mean square error ($RMSE_{CV}$%) from the LOOCV for predicted tree species richness and Shannon diversity index using LMMs and *k*-NN.

| Measures of tree species diversity | LMMs | | | | *k*-NN | | |
|---|---|---|---|---|---|---|---|
| | Predictor variables[a] | $R^2$ | $RMSE_{CV}$ | $RMSE_{CV}$ (%) | Predictor variables[a] | $RMSE_{CV}$ | $RMSE_{CV}$ (%) |
| Tree species richness | MaxF, PF20, TL0, TF2 | 0.46 | 2.12 | 40.7 | TF2, TF8, PF10, PL60, TL8, PL0, PF60, TF0, PF90, CVL, PF20 | 2.15 | 41.2 |
| Shannon diversity index | MaxF, PF10, TL0, TF2 | 0.39 | 0.45 | 39.1 | TF2, TF8, PF10, PL60, TL8, PL0, PF60, TF0, PF90, CVL, PF20 | 0.46 | 40.0 |

[a] PF10, PF20, PF60, PF90 = Percentiles of the first echo canopy heights for 10%, 20%, 60% and 90% (m).

PL60 = Percentiles of the last echo canopy heights for 60% (m);

TF2, TF8 = Canopy densities corresponding to the proportion of first echoes above fraction #2 and #8,

TL0, TL8, = Canopy densities corresponding to the proportion of last echoes above fraction #0 (1.3m), and #8,

MaxF = Maximum of the canopy height distributions of the first echoes,

CVL = Coefficient of variations of the last echo laser canopy heights.

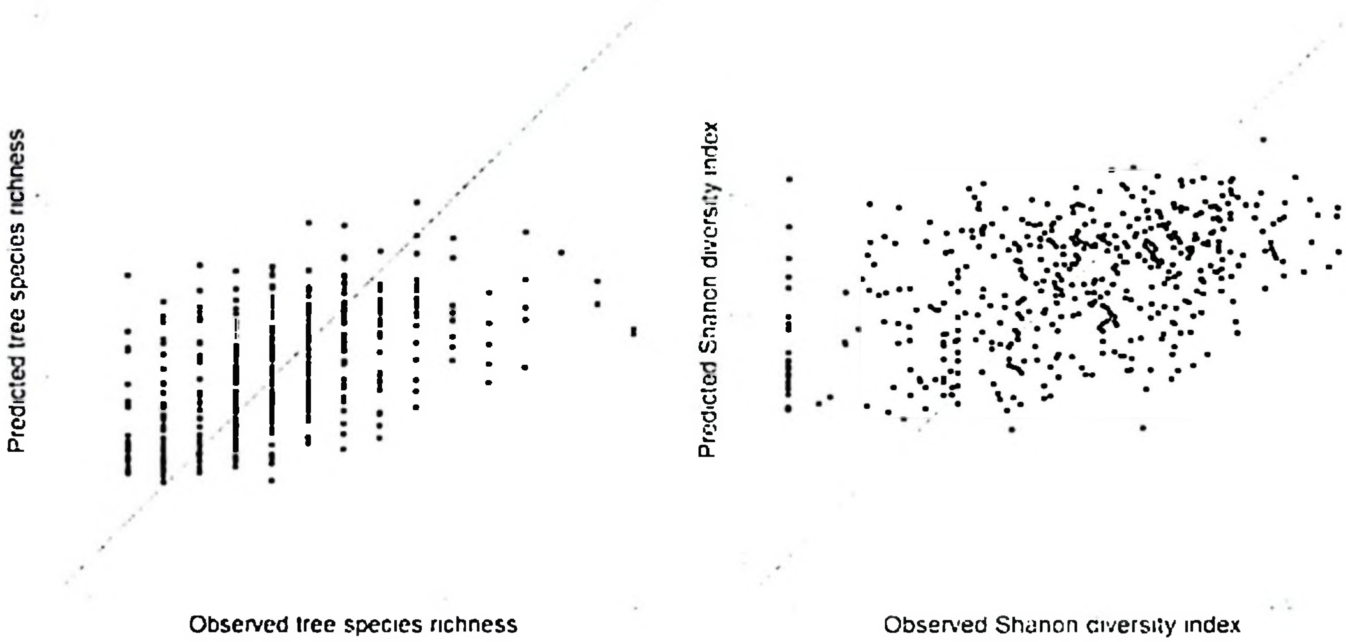



Fig. 5. Relationship between observed versus predicted richness (*left*) and Shannon diversity index (*right*) using LMMs.

*3.2. Non-parametric*

Multivariate predictions for both tree species richness and Shannon diversity index were performed using *k*-NN imputation models (Table 4). A total of eleven predictor variables were selected from the multivariate variable selection procedure. A comparison of RMSEcv% values calculated for tree species richness and Shannon diversity index using different number of neighbors is presented in Fig. 6. The imputation for both tree species richness and Shannon diversity index resulted in into lowest values of RMSEcv% when using imputation model with 10 neighbors. The RMSEcv% for Shannon diversity index was relatively low as compared to tree species richness. Comparing the two methods (i.e. LMMs and *k*-NN), the RMSEcv% for both tree species richness and Shannon diversity index were slightly lower for the LMMs compared to the *k*-NN. The validations plots for the *k*-NN imputations are presented in Fig. 7.

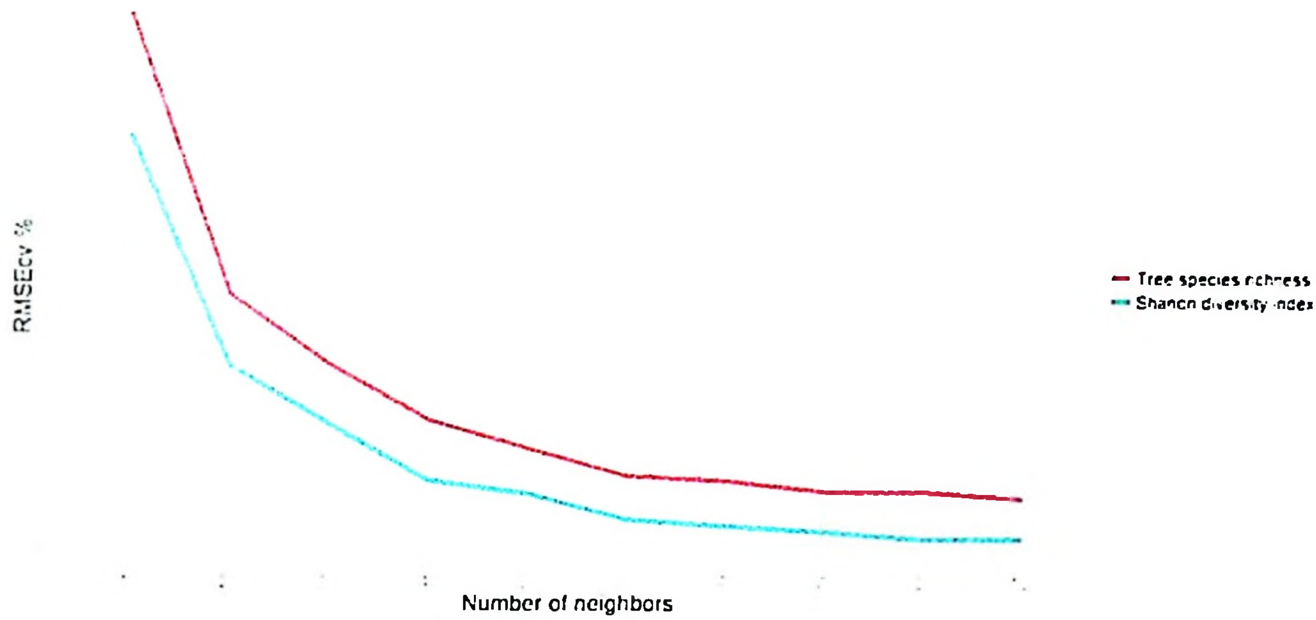



Fig. 6. Relative root mean square error from the LOOCV for different $k$-values (number of neighbors).

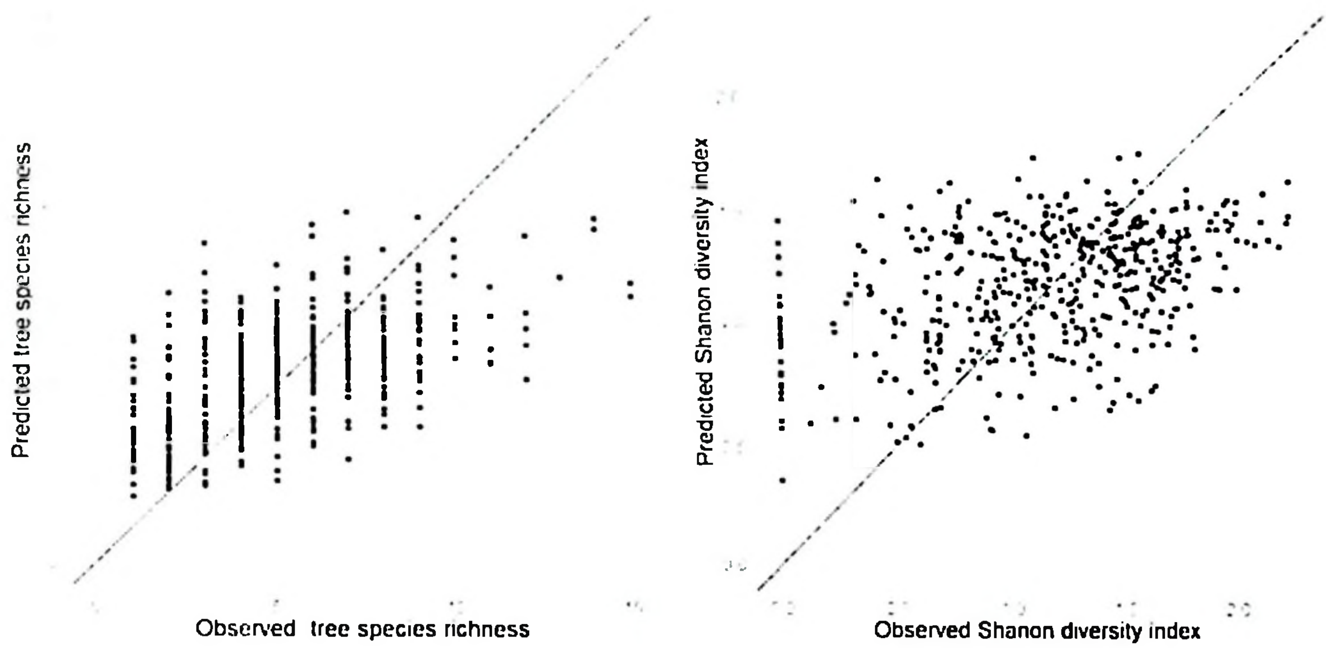



Fig. 7. Relationship between observed versus predicted richness (*left*) and Shannon diversity index (*right*) using $k$-NN imputations.

*3.3. Effect of vegetation types on prediction accuracy.*

The results suggest that vegetation types affect the relationship between the measures of tree species diversity and the ALS data. In the first approach, where the model fitted with dummy variables, indicated that for both tree species richness and Shannon diversity index, the parameter estimates for the dummy variables were significantly different from zero ($p < 0.05$) (Table 5). The standard error of parameter estimates for the LMMs with dummy variables were relatively low as compared to the LMMs without dummy variables. This is further shown by the results from the LOOCV where the RMSEcv% of the LMMs with dummy variables (Table 6) is relatively small as compared to the model without dummy variable. In the second approach where vegetation specific models were fitted, the influence of vegetation was proved to affect the prediction accuracy. Different predictor variables in each of the vegetation types were selected from the best subset procedure (Table 6). The variability explained by the vegetation specific LMMs ranged from 0.34 to 0.53 for tree species richness and from 0.32 to 0.47 for Shannon diversity index (Table 6). The results from the LOOCV indicated variation in prediction accuracy across the different vegetation types. Lowest RMSEcv% for predicting tree species richness across vegetation types was obtained in woodlands, while for Shannon diversity index the lowest value was obtained in forest (Table 6). Variations in prediction accuracy as measured by the RMSEcv% among the vegetation types were also observed when using *k*-NN (Table 7).

Table 5. Parameter estimates of LMMs for tree species richness and Shannon diversity index with dummy variables.

| Tree species richness | | | Shannon diversity index | | |
|---|---|---|---|---|---|
| Predictor[a] variables | Parameter estimates | Standard error | Predictor variables | Parameter estimates | Standard error |
| Intercept | 0.3377 | 0.0971 | Intercept | 0.2798 | 0.0909 |
| MaxF | 0.0315 | 0.0070 | MaxF | 0.0234 | 0.0063 |
| PF20 | -0.0760 | 0.0111 | PF10 | -0.0492 | 0.0120 |
| TL0 | -0.7082 | 0.1755 | TF2 | 1.3551 | 0.1859 |
| TF2 | 1.8926 | 0.1918 | TL0 | -0.6592 | 0.1701 |
| Forest | 0.4229 | 0.1183 | Forest | 0.2682 | 0.1117 |
| Woodlands | 0.4283 | 0.0796 | Woodlands | 0.2982 | 0.0748 |

[a] PF20 = Percentiles of the first echo canopy heights for 20% (m) ; TF2 = Canopy densities corresponding to the proportion of first echoes above fraction #2, TL0 = Canopy densities corresponding to the proportion of last echoes above fraction #0 (1.3m), MaxF = Maximum of the canopy height distributions of the first echoes; Forest and woodlands = Vegetation types according to MNRT (2011)

Table 6. Predictor variables, number of field sample plots (n), Pseudo-$R^2$ ($R^2$), absolute root mean square error ($RMSE_{cv}$), and relative root mean square error ($RMSE_{cv}$%) of LMMs for tree species richness and Shannon diversity index across different vegetation types.

| Measures of tree species diversity | Vegetation type[a] | Predictor variables[b] | n | $R^2$ | $RMSE_{cv}$ | $RMSE_{cv}$ (%) |
|---|---|---|---|---|---|---|
| Tree species richness | Woodlands | Maxf, PF20, TF3, TL0 | 387 | 0.41 | 2.08 | 38.9 |
| | Forest | PF20, TF9, PL50 | 40 | 0.34 | 2.79 | 41.8 |
| | Other cover types | MaxF, MeanF, TF1, TL0 | 57 | 0.36 | 1.84 | 57.0 |
| | All | MaxF, PF20, TL0, TF2, DUMMY | 484 | 0.45 | 2.10 | 40.2 |
| Shannon diversity index | Woodlands | CVF, TF4 | 387 | 0.32 | 0.42 | 35.0 |
| | Forest | PF40, TF8, MaxL, TL0 | 40 | 0.47 | 0.35 | 27.0 |
| | other cover types | MaxL, TF2 | 57 | 0.39 | 0.53 | 73.0 |
| | All | MaxF, PF10, TL0, TF2, DUMMY | 484 | 0.38 | 0.44 | 38.0 |

[a] Vegetation types according to MNRT (2011).
[b] PF10, PF20, PF40 = Percentiles of the first echo canopy heights for 10%, 20%, 40% (m).
PL50 = Percentiles of the last echo canopy heights for 50% (m).
TF1, TF2, TF3, TF4, TF8, TF9 = Canopy densities corresponding to the proportion of first echoes above fraction #1, # 2, #3, #4, #8, and #9.
TL0 = Canopy densities corresponding to the proportion of last echoes above fraction #0 (1.3m);
MeanF and MeanL = Arithmetic mean of first and last echo laser canopy heights (m).
MaxF and MaxL = Maximum of first and last echo laser canopy heights (m);
CVF and CVL = Coefficient of variations for the first and last echo respectively.
DUMMY= vegetation types (0 = Other cover types, 1 = Forest, 2 = Woodlands = 3).

Table 7. Predictor variables, number of field sample plots (n), absolute root mean square error (RMSEcv), and relative root mean square error (RMSEcv%) of the multivariate *k*-NN imputations for tree species richness and Shannon diversity index across different vegetation types.

| Measures of tree species diversity | Vegetation type [a] | Predictor variables | n | $RMSE_{CV}$ | $RMSE_{CV}$ (%) |
|---|---|---|---|---|---|
| Tree species richness | Woodlands | TF2, PL40, PF0, PF90, TF4 | 387 | 2.84 | 53.0 |
| | Forest | PL70, PF40, TL9, TL0, PF60, TF8, TF6, MaxL, TL2, PF90, TL3, PL10 | 40 | 2.99 | 44.8 |
| | Other cover types | TL0, TF1, PF80, PL60, TF8 | 57 | 2.93 | 90.7 |
| Shannon diversity index | Woodlands | TF2, PL40, PF0, PF90, TF4 | 387 | 0.59 | 49.2 |
| | Forest | PL70, PF40, TL9, TL0, PF60, TF8, TF6, MaxL, TL2, PF90, TL3, PL10 | 40 | 0.50 | 39.0 |
| | Other cover types | TL0, TF1, PF80, PL60, TF8 | 57 | 0.72 | 99.0 |

[a] Vegetation types according to MNRT (2011).

[b] PF10, PF40, PF90, PF80, PF90 = Percentiles of the first echo canopy heights for 10% , 40%, 90%, 80%, and 90% (m);
PL10, PL40, PL60, PL70 = Percentiles of the last echo canopy heights for 10%, 40%, 60, and 70% (m);
TF1, TF2, TF4, TF8, TF9 = Canopy densities corresponding to the proportion of first echoes above fraction #1, # 2 , # 3, #4, #8, and #9.;
TL0, TL2, TL3, and TL9 = Canopy densities corresponding to the proportion of last echoes above fraction #0 (1.3m), #2, #3, and #9;
MaxL = Maximum of last echo laser canopy heights (m).

## 4.0. Discussion.

The main objective of the study was to examine the usefulness of ALS data for modelling and predicting measures of tree species diversity in miombo woodlands of Tanzania. More specifically we tested and evaluated the performance of parametric and non-parametric methods for modelling and predicting tree species richness and Shannon diversity index using ALS data. Overall we found that ALS data can be used for modelling and predicting tree species richness and Shannon diversity index. Both tree species richness and Shannon diversity index are often used as measures of tree species diversity in remote sensing literature (e.g. Laurin et al., 2014; Leutner et al., 2012). However, to our understanding this is the first study to use ALS data for predicting tree species richness and Shannon diversity index in the miombo woodlands of Tanzania.

Based on our results from LMMs, more variations was explained by the ALS data when predicting tree species richness as compared to Shannon diversity index. Similar findings have been reported by Leutner et al. (2012) who assessed the potential of ALS and hyperspectral data for predicting plant species richness and Shannon diversity index in the temperate forests of Germany. The reason for this difference is unclear and opens questions for further investigations. However, irrespective of this difference, in an ecological context, Shannon diversity index is considered as a more appropriate measure of tree species diversity as compared to tree species richness, because it has the attributes of both richness and evenness. Besides that, Shannon diversity index is considered to be a better indicator for describing forest structure when using remotely sensed data as compared to tree species richness (Foody and Cutler, 2003). Use of Shannon diversity index in remote sensing based studies has also been suggested by others using both ALS and other remotely sensed data (e.g. Laurin et al., 2014; Oldeland et al., 2010)

Assessment of the performance of the two prediction methods, i.e. LMMs and the $k$-NN, has shown that both of the methods can reliably be used for predicting the measures of tree species diversity. The choice between them may therefore depend on the objective. For example, for making spatially consistent predictions of the multiple measures of tree species diversity, it is more reasonably to consider the use of $k$-NN. Furthermore, $k$-NN ($k$ =1) retains more of the original (co) variations than the LMMs. This is clearly seen in Fig. 8 where the predictions values from the $k$-NN imputations with $k$ = 1 were more frequently within the range of the observed values as compared to the LMMs and $k$-NN with $k$ = 10. This is in line with earlier studies that reported the strength of the $k$-NN over the parametric based methods (e.g. McRoberts et al., 2002; Moeur and Stage, 1995). LMM being a parametric method is more applicable if the interest is to examine the relationship between individual measures of tree diversity and the ALS data, which is important for deriving ecological interpretation of relationship between the measures of tree species diversity and the respective ALS metrics.

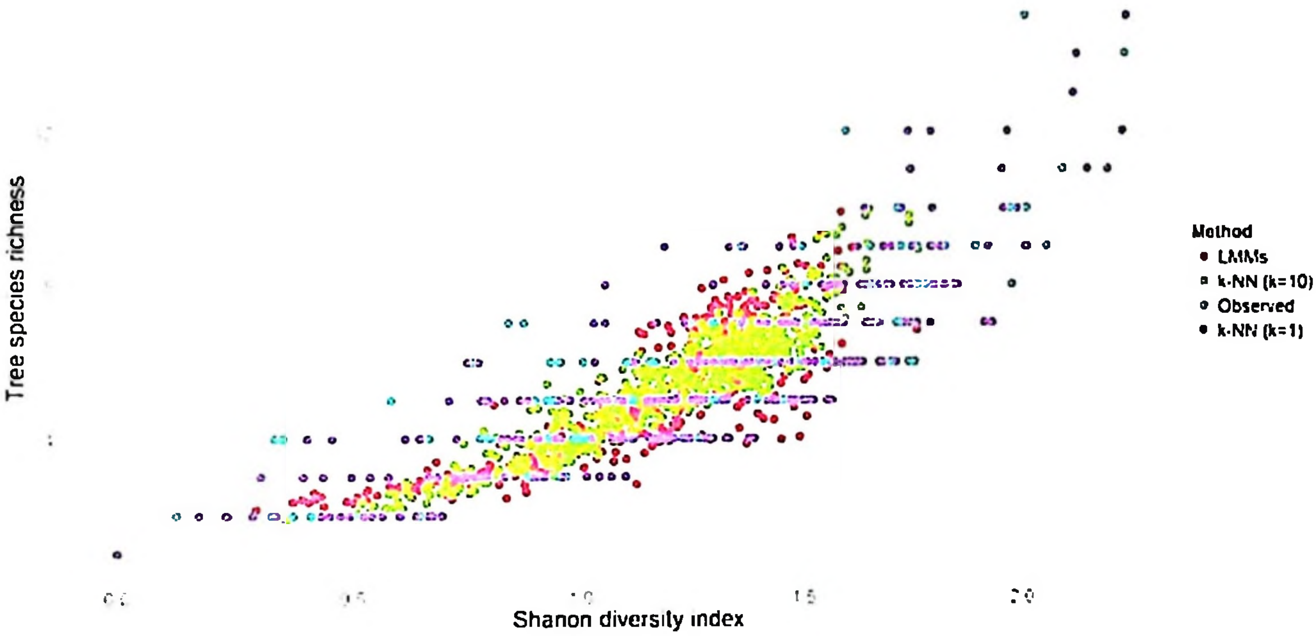



Fig 8. Observed and predicted values for tree species richness and Shannon diversity index.

The assessments of the effects of vegetation types on the prediction accuracy of both LMMs and *k*-NN have shown that vegetation types had significant impact on the prediction accuracy of the measures of tress species diversity when using ALS data. The parameter estimates of the dummy variables representing vegetation types were significantly different from zero when incorporated in the LMMs for predicting tree species richness and Shannon diversity index. This indicates that different vegetation types have different tree species diversity and spatial structure which entirely affect the relationship between the measures of tree species diversity and the ALS measurement. The importance of accounting for the effects of the vegetation types in the LMMs was also demonstrated by the results from LOOCV where the LMMs with dummy variables turned out to have lower RMSEcv% as compared to the LMMs without dummy variable. The variability imposed by different vegetation types was further indicated by different prediction accuracies when testing and evaluating both LMMs and *k*-NN across different vegetation types. As compared to LMMs the RMSEcv% values for the *k*-NN were relatively higher across different vegetation types. This might be attributed to the relatively low number of observations, since imputations methods are generally more sensitive to the number of observations as compared to parametric methods. The variability in prediction accuracy obtained across vegetation types, implies that it is important to control for the effect of vegetation types when making large scale estimation by post stratifying the area into different vegetation types.

Generally, the findings of our study in terms of model quality criterions such as $R^2$ and RMSEcv% are in accordance with other published studies that have attempted to use ALS data for predicting measures of tree species diversity in the tropical forests (Fricker et al., 2015; Wolf et al., 2012). For example, Wolf et al. (2012) reported $R^2$ of 0.48 when predicting tree species richness in the Neotropical forests of Panama. However, we should also mention that some of the studies from the tropical forests have reported relatively better results (e.g.

Hernández-Stefanoni et al., 2014; Laurin et al., 2014) than what we have obtained from this study. This is not surprising since in most cases the predictive ability of the remotely sensed data varies with the ecosystem and biogeographical context (Camathias et al., 2013). Furthermore, the differences in factors such as number of species, field design, data characteristics, and scales of reporting may attribute to the differences in prediction accuracy among the studies. It is therefore also important to mention some possible sources of errors that might have attributed to relatively low prediction accuracy in this study. The plot size used was relatively small compared to remote sensing based studies in ecology, where one hectare plot size has commonly been used (e.g. Asner and Mascaro, 2014; Asner et al., 2012; Mascaro et al., 2011). In ALS-based studies smaller plots have been reported to be a source of model errors due to the poor overlap between the field- and ALS-data. Thus, larger plots would most likely have helped in reducing the residual errors of the models. The effect of field plot size on ALS-based inventories is extensively analyzed in AGB related studies in the tropical forests (e.g. Mascaro et al., 2012; Mauya et al., 2015) and we would expect the same pattern in tree species diversity related studies. For example, Hernández-Stefanoni et al. (2014) compared the prediction accuracy of tree species richness at different plot size ranging from 400 $m^2$ to 2200 $m^2$ and indicated that the $R^2$ increased from 0.32 to 0.67. Another source of error that might have affected the accuracy of the model, is the NAFORMA subsampling procedure illustrated in Fig. 4. Since not all the trees within the plots were considered during measurement, we may expect a slight mismatch with what has been captured by the ALS data. Furthermore, with the NAFORMA subsampling procedure, it is possible that we have under-estimated tree species diversity, particularly when using Shannon diversity index, which theoretically assumes that all the species are represented in the sample.

While our study focused on the use of ALS alone, future studies should attempt to combine ALS with other remotely sensed data that have large spatial coverage such as hyperspectral images. This will likely also improve the prediction accuracy of the models, as the two types of sensors will complement each other, especially when modelling multi-layered forests (e.g. Dalponte et al., 2008; Laurin et al., 2014). Hyperspectral data are useful in providing detailed surface reflectance characteristics while ALS data adds detailed three-dimensional positioning information. Moreover, in this study, we focused on the commonly used ALS metrics that have widely been proved efficient in the prediction of various biophysical parameters including tree species diversity. However, we argue that further studies should be devoted in exploring other types of metrics such as texture based metrics that might improve the prediction accuracy of the models.

## 5.0. Conclusion

In this study the potential of using ALS data for prediction of measures of tree species diversity in the miombo woodlands of southern Tanzania has been analyzed. LMMs and *k*-NN imputations were tested and evaluated, and the results suggested that both approaches are promising tools for modelling and predicting measures of tree species diversity. The prediction accuracies of tree species richness and Shannon diversity index were affected by the differences in structural properties attributed by the vegetation types. Fusion of ALS data with other remote sensing techniques such as hyperspectral dataset is considered as an avenue for future research that might improve the prediction accuracy of trees species richness and Shannon diversity index.

## Acknowledgements

The financial support for this research was provided by the project entitled "Enhancing the Measuring, Reporting and Verification (MRV) of forests in Tanzania through the application of advanced remote sensing techniques". We are grateful to our field team in Tanzania for the field data collection and Terratec, Norway, for collecting and processing of the ALS data.

## References


Anon. 2012. TerraScan user's guide. Jyvaskyla, Finland: Terrasolid Oy. Available from http://www.terrasolid.com/download/tscan.pdf. 62. Accessed 23rd Aug 2014.

Asner, G.P., Mascaro, J., 2014. Mapping tropical forest carbon: Calibrating plot estimates to a simple LiDAR metric. Remote Sensing of Environment 140, 614-624.

Asner, G.P., Mascaro, J., Muller-Landau, H.C., Vieilledent, G., Vaudry, R., Rasamoelina, M., Hall, J.S., van Breugel, M., 2012. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia 168, 1147-1160.

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing 33, 111-118.

Barna, M., Bosela, M., 2015. Tree species diversity change in natural regeneration of a beech forest under different management. Forest Ecology and Management 342, 93-102.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F., 2012. Impacts of climate change on the future of biodiversity. Ecology letters 15, 365-377.

Borah, M., Das, D., Kalita, J., Deka Boruah, H.P., Phukan, B., Neog, B., 2015. Tree species composition, biomass and carbon stocks in two tropical forest of Assam. Biomass and Bioenergy 78, 25-35.

Boutin, S., Haughland, D.L., Schieck, J., Herbers, J., Bayne, E., 2009. A new approach to forest biodiversity monitoring in Canada. Forest Ecology and Management 258, S168-S175.

Camathias, L., Bergamini, A., Küchler, M., Stofer, S., Baltensweiler, A., 2013. High-resolution remote sensing data improves models of species richness. Applied Vegetation Science 16, 539-551.

Ceballos, A., Hernández, J., Corvalán, P., Galleguillos, M., 2015. Comparison of Airborne LiDAR and Satellite Hyperspectral Remote Sensing to Estimate Vascular Plant

Richness in Deciduous Mediterranean Forests of Central Chile. Remote Sensing 7, 2692-2714.

Chandra, A., Idrisova, A., 2011. Convention on Biological Diversity: a review of national challenges and opportunities for implementation. Biodiversity and Conservation 20, 3295-3316.

Chao, A., Shen, T.-J., 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and ecological statistics 10, 429-443.

Chiarucci, A., Palmer, M., 2006. The inventory and estimation of plant species richness. Encyclopedia of life support systems, developed under the auspices of the UNESCO. Encyclopedia of Life Support Systems Publishers, Oxford.

Chidumayo, E.N., 2013. Forest degradation and recovery in a miombo woodland landscape in Zambia: 22 years of observations on permanent sample plots. Forest Ecology and Management 291, 154-161.

Chirici, G., McRoberts, R.E., Winter, S., Bertini, R., Brändli, U.-B., Asensio, I.A., Bastrup-Birk, A., Rondeux, J., Barsoum, N., Marchetti, M., 2012. National forest inventory contributions to forest biodiversity monitoring. Forest Science 58, 257-268.

Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society B: Biological Sciences 345, 101-118.

Crookston, N.L., Finley, A.O., 2008. yaimpute: An r package for knn imputation. Journal of Statistical Software 23, 1-16.

Dalponte, M., Bruzzone, L., Gianelle, D., 2008. Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. Geoscience and Remote Sensing, IEEE Transactions on 46, 1416-1427.

Drew, C.A., Wiersma, Y.F., Huettmann, F., 2010. Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer Science & Business Media.

Ehara, M., Hyakumura, K., Yokota, Y., 2014. REDD+ initiatives for safeguarding biodiversity and ecosystem services: harmonizing sets of standards for national application. Journal of Forest Research 19, 427-436.

Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T., 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research 24, 235-246.

Foody, G.M., Cutler, M.E., 2003. Tree biodiversity in protected and logged Bornean tropical rain forests and its measurement by satellite remote sensing. Journal of Biogeography 30, 1053-1066.

Fricker, G.A., Wolf, J.A., Saatchi, S.S., Gillespie, T.W., 2015. Predicting spatial variations of tree species richness in tropical forests from high resolution remote sensing. Ecological Applications. Ecological Applications., http://dx.doi.org/10.1890/14-1593.1

Gardner, T.A., Barlow, J., Chazdon, R., Ewers, R.M., Harvey, C.A., Peres, C.A., Sodhi, N.S., 2009. Prospects for tropical forest biodiversity in a human-modified world. Ecology Letters 12, 561-582.

Gardner, T.A., Burgess, N.D., Aguilar-Amuchastegui, N., Barlow, J., Berenguer, E., Clements, T., Danielsen, F., Ferreira, J., Foden, W., Kapos, V., 2012. A framework for integrating biodiversity concerns into national REDD+ programmes. Biological Conservation 154, 61-71.

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Holm, S., Ørka, H.O., Astrup, R., 2012. Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment 123, 443-456.

Hernández-Stefanoni, J.L., Dupuy, J.M., Johnson, K.D., Birdsey, R., Tun-Dzul, F., Peduzzi, A., Caamal-Sosa, J.P., Sánchez-Santos, G., López-Merlín, D., 2014. Improving Species Diversity and Biomass Estimates of Tropical Dry Forests Using Airborne LiDAR. Remote Sensing 6, 4741-4763.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., Falkowski, M.J., 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment 112, 2232-2245.

Imai, N., Tanaka, A., Samejima, H., Sugau, J.B., Pereira, J.T., Titin, J., Kurniawan, Y., Kitayama, K., 2014. Tree community composition as an indicator in biodiversity monitoring of REDD+. Forest Ecology and Management 313, 169-179.

Kindt, R., Coe, R., 2005. Tree diversity analysis: a manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre.

Latifi, H., Fassnacht, F.E., Hartig, F., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2015. Stratified aboveground forest biomass estimation by remote sensing data. International Journal of Applied Earth Observation and Geoinformation 38, 229-241.

Laurin, G.V., Chan, J.C.-W., Chen, Q., Lindsell, J.A., Coomes, D.A., Guerriero, L., Del Frate, F., Miglietta, F., Valentini, R., 2014. Biodiversity mapping in a tropical west african forest with airborne hyperspectral data. PloS one 9, e97910.

LDC, 2014. Liwale Distric Socio-Economic Profile, Unpublished Report, p. 56.

Leutner, B.F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., Wegmann, M., 2012. Modelling forest α-diversity and floristic composition—On the added value of LiDAR plus hyperspectral remote sensing. Remote Sensing 4, 2818-2845.

Lindberg, E., Roberge, J.-M., Johansson, T., Hjältén, J., 2015. Can Airborne Laser Scanning (ALS) and Forest Estimates Derived from Satellite Images Be Used to Predict Abundance and Species Richness of Birds and Beetles in Boreal Forest? Remote Sensing 7, 4233-4252.

Magurran, A.E., 1988. Ecological diversity and its measurement. Princeton university press Princeton.

Mascaro, J., Asner, G.P., Dent, D.H., DeWalt, S.J., Denslow, J.S., 2012. Scale-dependence of aboveground carbon accumulation in secondary forests of Panama: A test of the intermediate peak hypothesis. Forest Ecology and Management 276, 62-70.

Mascaro, J., Detto, M., Asner, G.P., Muller-Landau, H.C., 2011. Evaluating uncertainty in mapping forest carbon with airborne LiDAR. Remote Sensing of Environment 115, 3770-3774.

Matthews, R.B., van Noordwijk, M., 2014. From euphoria to reality on efforts to reduce emissions from deforestation and forest degradation (REDD+). Mitigation and Adaptation Strategies for Global Change 19, 615-620.

Mauya, E.W., Hansen E. H., Gobakken, T., Bollandsås, O.M., Malimbwi, R., Næsset, E., 2015. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. Carbon balance and management 10, 1-14.

McRoberts, R.E., 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management 272, 3-12.

McRoberts, R.E., Meneguzzo, D.M., 2005. Estimating tree species richness from forest inventory plot data. Proceedings of the seventh annual forest inventory and analysis symposium, Portland, ME, pp. 3-6.

McRoberts, R.E., Nelson, M.D., 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.

McRoberts, R.E., Næsset, E., Gobakken, T., 2015. Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. Remote Sensing of Environment 163, 13-22.

MNRT. NAFORMA Field Manual - Biophysical. http:// http://www.fao.org/forestry/23484-05b4a32815ecc769685b21b03be44ea77.pdf (2011). Accessed 23rd Feb 2014.

MNRT. 2015. National Forest Resources Monitoring and Assessment of Tanzania Mainland (NAFORMA). Main Results. http://www.fao.org/forestry/43612-09cf2f02c20b55c1c00569e679197dcde.pdf. Accessed 17th Aug 2015.

Moeur, M., Stage, A.R., 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science 41, 337-359.

Mooney, H.A., Cropper, A., Reid, W., 2004. The millennium ecosystem assessment: what is it all about? Trends in Ecology & Evolution 19, 221-224.

Müller, J., Brandl, R., 2009. Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. Journal of Applied Ecology 46, 897-905.

Müller, J., Vierling, K., 2014. Assessing biodiversity by airborne laser scanning In: Maltamo M, Næsset E, Vauhkonen J, editors. Forestry applications of airborne laser scanning – concepts and case studies. Dordrecht, Netherlands:Springer, pp. 357-374.

Nadeau, M.B., Sullivan, T.P., 2015. Relationships between Plant Biodiversity and Soil Fertility in a Mature Tropical Forest, Costa Rica. International Journal of Forestry Research 2015.

Næsset, E., 2004. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research 19, 164-179.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity 41, 673-690.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Stevens, M., Wagner, H., 2011. Vegan: community ecology package version 2.0-2. R CRAN package.

Oldeland, J., Wesuls, D., Rocchini, D., Schmidt, M., Jürgens, N., 2010. Does using species abundance data improve estimates of species diversity from remotely sensed spectral heterogeneity? Ecological Indicators 10, 390-396.

Pereira, H.M., Cooper, H.D., 2006. Towards the global monitoring of biodiversity change. Trends in Ecology & Evolution 21, 123-129.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G., Jongman, R., Scholes, R., Bruford, M., Brummitt, N., Butchart, S., Cardoso, A., 2013. Essential biodiversity variables. Science 339, 277-278.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S-PLUS. Springer.

Roe, S., Streck, C., Pritchard, L., Costenbader, J., 2013. Safeguards in REDD+ and forest carbon standards: A review of social, environmental and procedural concepts and application. Climate Focus.

Schmidt, M., Veldkamp, E., Corre, M.D., 2015. Tree species diversity effects on productivity, soil nutrient availability and nutrient response efficiency in a temperate deciduous forest. Forest Ecology and Management 338, 114-123.

Shirima, D.D., Totland, Ø., Munishi, P.K.T., Moe, S.R., 2015. Relationships between tree species richness, evenness and aboveground carbon storage in montane forests and miombo woodlands of Tanzania. Basic and Applied Ecology 16, 239-249.

Simonson, W.D., Allen, H.D., Coomes, D.A., 2012. Use of an airborne lidar system to model plant species composition and diversity of Mediterranean oak forests. Conservation Biology 26, 840-850.

Snowdon, P., 1991. A ratio estimator for bias correction in logarithmic regressions. Canadian Journal of Forest Research 21, 720-724.

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., Blair, J.B., 2011. Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. Remote Sensing of Environment 115, 2917-2930.

Team, R.C., 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. Open access available at: http://cran. r-project. org.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H.M., Chamuya, N., Zahabu, E., Otieno, J., 2014. A sampling design for a large area forest inventory: case Tanzania. Canadian Journal of Forest Research 44, 931-948.

Vanderhaegen, K., Verbist, B., Hundera, K., Muys, B., 2015. REALU vs. REDD+: Carbon and biodiversity in the Afromontane landscapes of SW Ethiopia. Forest Ecology and Management 343, 22-33.

Vierling, K., Bässler, C., Brandl, R., Vierling, L., Weiß, I., Müller, J., 2011. Spinning a laser web: predicting spider distributions using LiDAR. Ecological Applications 21, 577-588.

Vogeler, J.C., Hudak, A.T., Vierling, L.A., Evans, J., Green, P., Vierling, K.T., 2014. Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. Remote Sensing of Environment 147, 13-22.

Winter, S., McRoberts, R.E., Chirici, G., Bastrup-Birk, A., Rondeux, J., Brändli, U.-B., Nilsen, J.-E.Ø., Marchetti, M., 2011. The need for harmonized estimates of forest biodiversity indicators. National Forest Inventories: Contributions to Forest Biodiversity Assessments. Springer. pp. 1-23.

Wolf, J.A., Fricker, G.A., Meyer, V., Hubbell, S.P., Gillespie, T.W., Saatchi, S.S., 2012. Plant species richness is associated with canopy height and topography in a neotropical forest. Remote Sensing 4, 4010-4021.

**Appendix. Species accumulation curves.**

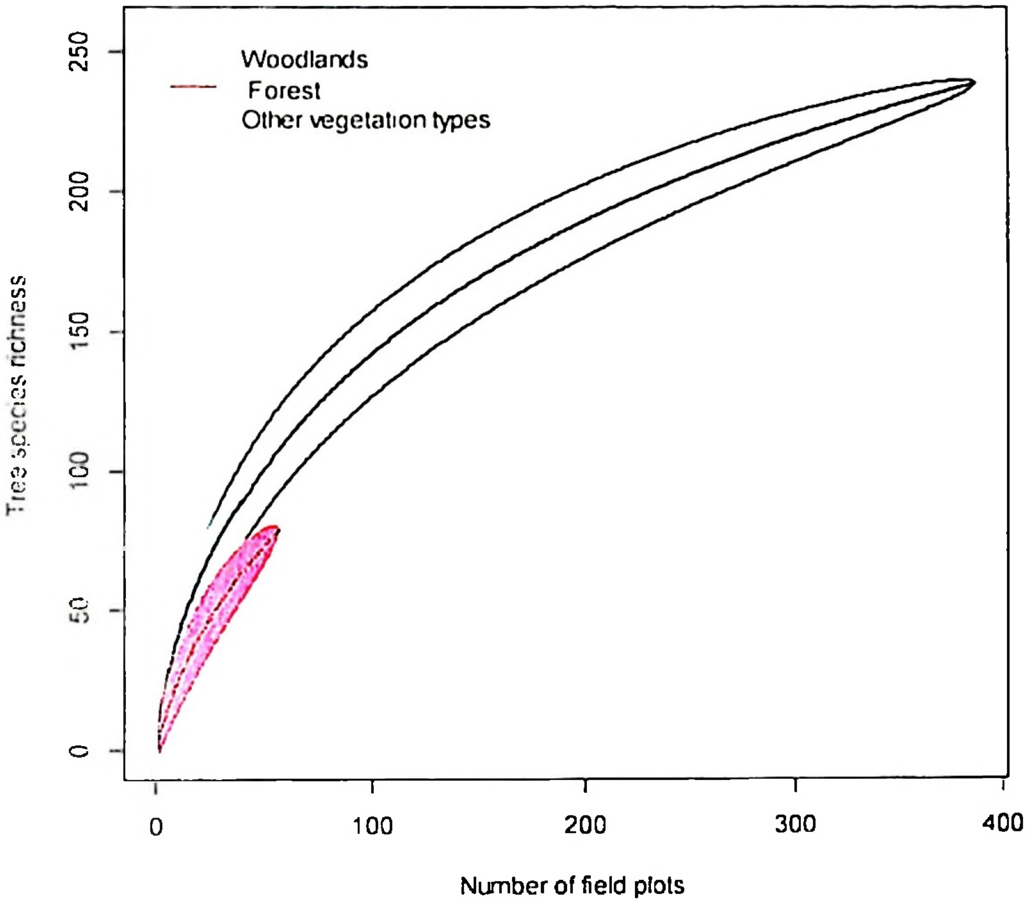



Fig. A1. Species accumulation curves for different vegetation types.

W

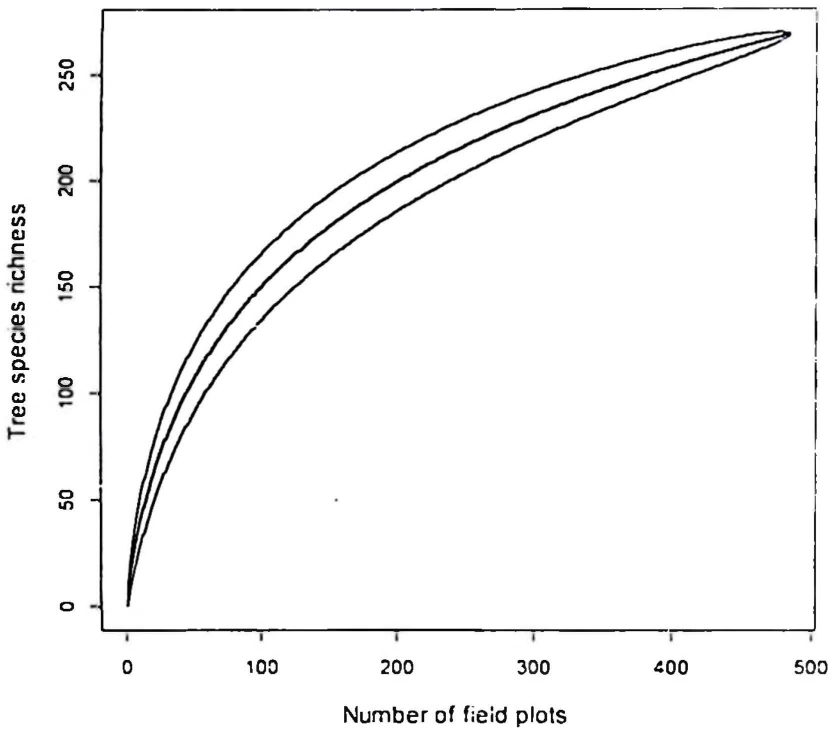



Fig. A2. Species accumulation curve for all field sample plots pulled together.