ESTIMATING FOREST STRUCTURAL CHARACTERISTICS USING THE AIRBORNE LIDAR SCANNING SYSTEM AND A NEAR-REAL TIME PROFILING LASER SYSTEM

A Dissertation

by

KAIGUANG ZHAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2008

Major Subject: Forestry

ESTIMATING FOREST STRUCTURAL CHARACTERISTICS USING THE AIRBORNE LIDAR SCANNING SYSTEM AND A NEAR-REAL TIME PROFILING LASER SYSTEM

A Dissertation

by

KAIGUANG ZHAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Sorin C. Popescu
Committee Members,	Ross F. Nelson
	X. Ben Wu
	Marian Eriksson
Head of Department,	Steven G. Whisenant

August 2008

Major Subject: Forestry

ABSTRACT

Estimating Forest Structural Characteristics Using the Airborne Lidar Scanning System and a Near-Real Time Profiling Laser System. (August 2008)
Kaiguang Zhao, B.S.; M.S., Beijing Normal University
Chair of Advisory Committee: Dr. Sorin C. Popescu

LiDAR (Light Detection and Ranging) directly measures canopy vertical structures, and provides an effective remote sensing solution to accurate and spatially-explicit mapping of forest characteristics, such as canopy height and Leaf Area Index. However, many factors, such as large data volume and high costs for data acquisition, precludes the operational and practical use of most currently available LiDARs for frequent and large-scale mapping. At the same time, a growing need is arising for real-time remote sensing platforms, e.g., to provide timely information for urgent applications. This study aims to develop an airborne profiling LiDAR system, featured with on-the-fly data processing, for near real- or real- time forest inventory. The development of such a system involves implementing the on-board data processing and analysis as well as building useful regression-based models to relate LiDAR measurements with forest biophysical parameters.

This work established a paradigm for an on-the-fly airborne profiling LiDAR system to inventory regional forest resources in real- or near real- time. The system was developed based on an existing portable airborne laser system (PALS) that has been previously assembled at NASA by Dr. Ross Nelson. Key issues in automating PALS as an on-the-fly system were addressed, including the design of an archetype for the system workflow, the development of efficient and robust algorithms for automatic data processing and analysis, the development of effective regression models to predict forest biophysical parameters from LiDAR measurements, and the implementation of an integrated software package to incorporate all the above development.

This work exploited the untouched potential of airborne laser profilers for real-time forest inventory, and therefore, documented an initial step toward developing airborne-laser-based, on-the-fly, real-time, forest inventory systems. Results from this work demonstrated the utility and effectiveness of airborne scanning or profiling laser systems for remotely measuring various forest structural attributes at a range of scales, i.e., from individual tree, plot, stand and up to regional levels. The system not only provides a regional assessment tool, one that can be used to repeatedly, remotely measure hundreds or thousands of square kilometers with little/no analyst interaction or interpretation, but also serves as a paradigm for future efforts in building more advanced airborne laser systems such as real-time laser scanners.

DEDICATION

To my parents and sisters, and to the great mentors who shaped me as am I today.

ACKNOWLEDGEMENTS

This dissertation could not have been made possible without the support of many individuals. First of all, I am privileged to have Dr. Sorin Popescu as the chair of my committee. His constant assistance and inspiration started from a correspondence email back four years ago when I applied for the graduate school, and since then, his mentoring has been "enriching" me in all aspects of my study and life. I am also lucky enough to have Dr. Ross Nelson, Dr. Ben Wu and Dr. Marian Eriksson serve on my advisory committee. Dr. Ross Nelson affected me so much with his greatest patience and passions, leaving too much for me to learn from him. Dr. Ben Wu and Dr. Marian Eriksson touched me with their respective expertise, and I gained a lot in building my knowledge, especially in ecology and programming. I express my sincerest gratitude to each one of my committee for their guidance and support in both professional and personal aspects.

I acknowledge the support by the Texas Forest Service; in particular, I thank Curt Stripling and other personnel of the Texas Forest Service for field data collection that was used in this research. I also thank the following professors for their help at the different stages of my Ph.D. study: Drs. Rusty Feagin, Ping Yang, Raghavan Srinivasan, Mark Tjoelker, and Hongxing Liu.

Many friends "colorized" my "monotone" life in College Station. Their companionship enlivened my academic study. I thank all of them: Xuesong Zhang, Muge Mutlu, Alica Griffin, Jared Stueky, Yue Chen, Yuanchang Xie, Feng Liu, E Bai, Xiaojing Duan, Daniel Grimes, Shawn Grimes, Mom Zsuzsa, and Mr. and Mrs. Wada.

My unreserved thanks go to all my family in China: my parents Zhao Qingan, Sun Peiai, and my three sisters, Zhao Kaie, Zhao Kaijing, Zhao Kaimei I owe special thanks to my eldest sister Zhao Kaie who sacrificed her education opportunity to fulfill mine, and if I could split it, this degree also belongs to her.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

Page

Page

Page

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

1.1 Remote Sensing for Forest Resources Inventory

Forests are some of the most important natural resources for human beings, and the importance lies in the roles of forests not only as habitats for wildlife, sources of wood for timber industries, resources for human recreation, but also as a regulating factor of local or global environments. Forests, for example, serve as a major reservoir of carbon terrestrial ecosystems by sequestering atmospheric carbon dioxide, which is a complex process coupled with other factors such as climate, land use, and anthropogenic activities to dominate the global environmental changes (Houghton, 2007). Accurate inventory of forest resources, therefore, is crucial to better characterize the functions of forests and understand the predicted responses of forests to climatic variations. Up to now, many efforts have been dedicated to the survey of forest resources at various scales (Parresol, 1999; Chan et al., 2003). The Forest Inventory and Analysis (FIA) program of the USDA Forest Service, for example, offers one of the most intensive surveys designed to monitor the USA's forests with a continuous collection starting from 1930 (Alerich et al., 2004). The FIA data could be used to monitor and project the status and dynamics of forests at regional or national scales.

Although field collection of forest inventory data provides detailed information on forest structural characteristics that can be used for further studies, fieldwork is often labor-intensive, time-demanding, prone to errors and difficult to repeat. Moreover, field measurements are collected at isolated sites or plots, thus failing to provide spatially

---

This thesis follows the style of *Remote Sensing of Environment*.

continuous observation over large areas. As an alternative, remote sensing has brought new and efficient technologies to remotely collect the earth surface information and map natural resources with spatially continuous and temporally repeatable coverage. At the heart of remote sensing is the use of electromagnetic radiation to "sense" attributes of targets without direct contact with the targets (Liang, 2004; Turner et al., 2004).

Conventional remote sensing approaches such as visible to mid-infrared optical sensors prove successful in numerous applications relevant to natural resources inventory (Carter, 1991; Wulder, 1998; Li et al, 2000; Cohen and Goward, 2004). Retrieval of canopy biophysical parameters from multi- or hyper- spectral imagery was made possible either by statistical- or physical-based methods, but only with partial success over moderate to high biomass forests. Statistical methods are usually study- or site- specific. Physical-based retrieval schemes usually rely on forward models that describe the remote sensing processes. Ironically, more often than not, the more complicated the forward modeling, the better the model mimics the real imaging processes, but the less accurate the information retrieved by the forward-model-based inversion algorithm (Liang, 2004). Furthermore, the representation of spectral measurements from conventional optical sensors is presented as multiple 2D images, and optical responses are insensitive to dense canopies; therefore, forest vertical structures are difficult to characterize by only using multi-spectral images. On the other hand, although photogrammetry is a well-established technique capable of deriving 3D information on the observed top surfaces, it fails to detect signals from the ground under relatively dense canopy conditions due to less opening (Popescu, 2002).

A breakthough in forestry remote sensing is LIght Detection And Ranging (LiDAR), a recently emerged technique to characterize forest vertical structures with its capability of directly measuring forest 3D structures (Lefsky et al., 2002). In principle, LiDAR transmits pulsed lasers to accurately pinpoint targets by timing the roundtrip between the pulse and the intercept targets and multiplying the time of flight with the speed of light to get a range measurement. If the intercept target has a multilayer structure, the reflected energy will be a short-lived yet continuous waveform where

multiple peaks may be present, and by appropriately identifying these peaks through a triggering threshold of magnitude, multiple objects can be detected in the illuminating path of the pulse (Lefsky et al., 2002; Lim et al., 2003; Popescu, 2002). Such a case with multiple laser returns is illustrated in Figure 1.1 for a tree, i.e., a typical example of multi-layered feature.

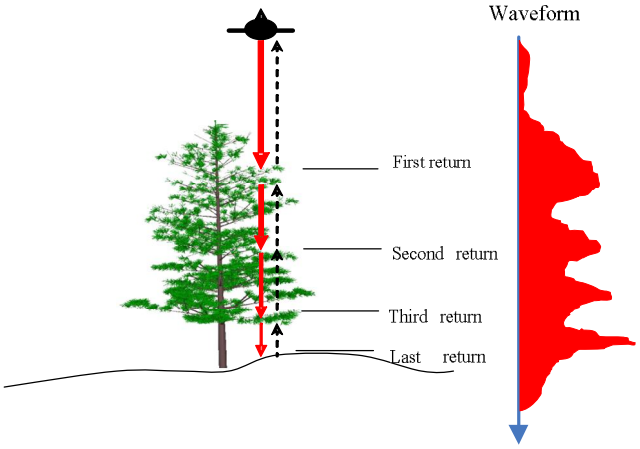


Figure 1.1. An illustration of multiple returns from a single laser pulse where the decreasing width of red line indicates the attenuation of laser energy as it propagates: discrete returns (the left) and waveform (the right)

1.2 LiDAR Background

LiDAR systems are in use to meet needs of different disciplines, e.g., for the purposes of mapping depth of water bodies in bathymetry, and measuring aerosol in the atmospheric sciences. For terrestrial applications such as forest inventory, two types of LiDAR systems are often differentiated, namely, waveform or discrete-return LiDAR (Lim et al., 2003). The former digitizes and records reflected energy continuously with a fine time resolution; however, the latter only records several discrete locations at which the magnitudes are above a preset threshold to trigger recording (Axelsson, 1999). Besides, LiDARs can also be categorized in terms of other standards. An incomplete summary of types of LiDARs according to different criteria is listed in Table 1.1. The

same LiDAR system can be classified differently with respect to various criteria. For example, the system developed by Gutierrez et al. (2005) can be termed as an airborne, small-footprint, waveform, profiling LiDAR for terrestrial applications. But it is noted that not every combination of these types is currently available; for example, no space-borne small-footprint discrete-return LiDAR is available at the time of writing.

Table 1.1. Classification of LiDAR systems according to different criteria

Platform	Footprint Size	Digitization Mode	Operating Mode	Applications
ground-based, airborne, space-borne	small-footprint, large-footprint	discrete-return, waveform	profiling, scanning	oceanic, terrestrial, atmospheric

A LiDAR system operated in a profiling mode usually emits laser pulses at near-nadir direction, and therefore, it collects ranging measurements only along the linear transects that lie directly under flight lines. Figure 2 shows a typical example of a forest canopy profile obtained by profiling LiDARs along the ground track. Figure 1.2a suggests that two profiles are involved to characterize stand structures: the top profile depicts the envelop of the top canopy surface, and the bottom one describes the terrain topography. In fact, the idea to produce a stand profile using accurate instruments was presented by Hugershoff as early as in 1939. The introduction of LiDAR into terrain applications were evolved from the oceangraphic applications of LiDAR (Hyyppä and Hallikainen, 1996). In 1984, Nelson proposed to use a profiling laser system to obtain forest stand profiles for the retrieval of stand characteristics, and found that the elements of the stand profile are linearly related to crown closure and may be used to assess tree height.

Airborne laser scanners (ALS) collect data over strips of forested areas by oscillating laser pulses, thus, enabling mapping the full extent of a study area with multiple swaths (Figure 1.2b). Modern ALS systems are often equipped with a Differential GPS to record the trajectory of the platform and with an Inertial Measurement Unit (IMU) to track the altitude of the platform (yaw, roll, and pitch);

therefore, every laser return can be geo-referenced (Ackermann, 1999). Data acquired by a discrete-return ALS consist of a cloud of laser returns with each corresponding to a geographical coordinate tripe of (x, y, z) and often tagged with auxiliary information such as intensity, acquisition time, and scanning angle. Most of the commercially available LiDARs belong to ALS, and the use of ALS data for forest inventory at different scales has been extensively investigated. Previous studies suggest that ALS holds immense potentials for spatially-explicit mapping of forest resources (Maltamo et al., 2000; Zimble et al., 2003; Riaño et al., 2004; Næsset, 2004; Andersen et al., 2005; Maltamo et al., 2006; Mutlu et al., 2008).

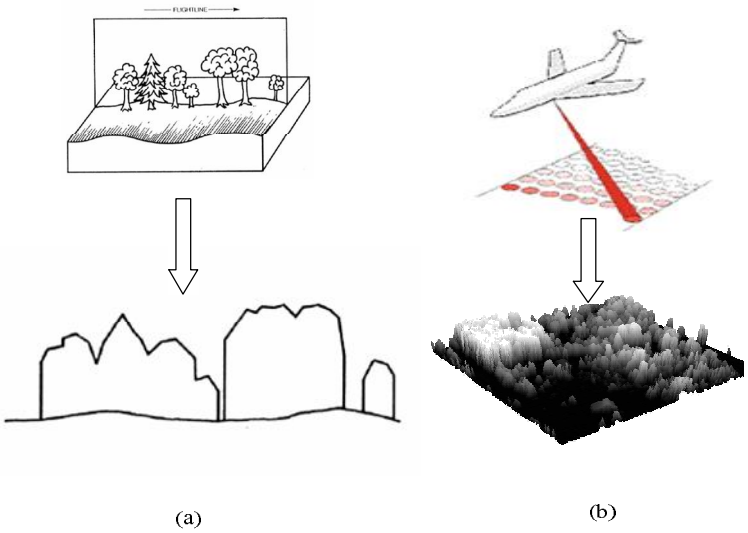


Figure 1.2. Schematics of two types of LiDAR systems for measuring forest stands: (a) A profiling LiDAR, and (b) a scanning LiDAR

1.3 Literature Review

1.3.1 Profiling LiDAR

Airborne laser profilers were among the first generation of LiDAR systems for forestry applications. Nelson et al. (1984) proposed the use of an airborne laser profiler to acquire forest stand profiles for the retrieval of stand characteristics over a oak-

hickory forest, and found that laser-based estimates of canopy height is underestimated by an average of 0.6 m in reference to photogrammetric estimates, but with the laser estimates being more accurate. They also reported that the penetration rate of laser pulses into canopy is highly correlated with canopy closure. Schreier et al. (1985) employed a laser system that records both ranging and magnitude of each pulse and produces a pulse density of 6.7 pulses per square meter over forests in northeastern Ontario, Canada; based on laser height, laser reflection and reflection variability parameters, they applied a semi-automated classification approach to distinguish conifer from broadleaf forest types. Their research also indicated that the laser system is able to trace individual trees and terrain height profiles, and also to obtain accurate estimates of tree heights. Maclean and Krabill (1986) found that laser-based canopy profile area is a very good predictor of total gross-merchantable timber and that strata-specific predictions improved the strength of correlation. They also suggested that the canopy profile area could be used to estimate total biomass or total woody biomass.

The potential use of laser profiling data for estimating forest biomass and the repeatability of the laser observations were explored in Nelson et al. (1988a). In this study, they found that 3-6% laser estimates of volume and biomass varied between comparable sections of different overpasses along the same flight line. Their study showed there are 7% and 8% differences of biomass and volume between laser estimates and ground measurements while the differences between the two sets of laser estimate are 0.9% and 1.2% for biomass and volume, respectively. They concluded that the advantages of using an airborne profiling laser to collect data over forests are twofold: first, canopy height data can be collected quickly along transects hundreds of miles long; second, laser data can be used for extending a limited effort of ground sampling to areas that may not be easily accessible by ground inventory crews. In another study, Nelson et al. (1988b) estimated biomass and volume by building different regression models and using six laser height variables as potential predictors. The best fitted models explained between 53% and 65% of variability in the ground measurements of forest biomass and

volume. Study results also indicated that species stratification did not consistently improve regression relationships for four southern pine species.

Ritchie et al. (1992) measured canopy cover and distribution in two rangelands of south Texas with an airborne profiling laser; a good agreement with a R^2 of 0.89 was observed between the ground-measured canopy cover and laser estimates for 18 30.5-m segments. They stated that airborne laser data can provide quick and reliable measurements of canopy cover and distributions over large rangeland areas. Using high-resolution laser profiler data, Nelson et al. (1997) estimated tree heights, canopy density, basal area, and biomass over tropical forests without relying on LiDAR-ground transect co-location; instead, they used the simulated canopy laser measurements to develop multiple regression models for predicting for basal area, volume, and biomass.

More recently, an inexpensive, lightweight, airborne-profiling system (PALS) has been assembled from off-the-shelf, commercially available components (Nelson et al., 2003). The system, with a total cost of US$30,000, includes a laser transmitter/receiver, a differential GPS, a charge-coupled device video camera, a laptop computer, and an integration software package (LabVIEW) to coordinate and store the different data streams, any hardware component of which may be replaced to meet specialized needs. PALS is designed to fly aboard small helicopters and single or twin-engine high wing aircraft without airframe modification. It has a data rate up to 2000 hz with an operational envelope up to 300 m above ground. PALS measures both range and amplitude along flight transects; it was originally designed to collect only first returns, but a later upgrade enabled it to toggle between measurements of first and last returns (Nelson et al., 2005). This economical laser system was reported to be capable of conducting regional or subcontinental forest inventory worldwide. In the subsequent studies, Nelson et al. (2004) used the PALS data acquired along transects spaced 4 km apart to measure biomass and carbon in the state of Delaware, USA. Four linear models were developed to regress the merchantable volume and total aboveground dry biomass against laser-based predictors. The estimates at the level of transect were extended to the areal estimates based on the line intercept sampling techniques. The laser-based

estimates of merchantable volume fell within 21% of US forest service estimates at the county level, and 1% statewide. Estimates of above-ground dry biomass were within 22% of USFS estimates at the county level and 16% statewide. They also used the profiling measurements to estimate the area percentage of impervious and open water area for the three counties of Delaware, and the results were comparable to those from other sources. Their study indicated that the line intercept sampling techniques can be used in conjunction with PALS to investigate multiple resources regionally. Nelson et al. (2005) located and estimated the extent of potential Delmarva fox squirrel (DFS) habitats using the same profiling dataset of Delaware. Since this particular dataset of profiling measurements contain little or no understory vertical structure, the identified 53 sites are potential but not actual DFS habitat; however, this helped screen the extensive unsuitable areas. In addition, they estimated the forest percentage according to height/canopy closures levels at both the state and county levels; the average within-patch crossing distance and average between-patch distances of the landscape were also reported. Skowronski et al. (2007) analyzed profiling LiDAR data over the Pinelands of New Jersey for measuring canopy structure and fuel loads, and their results indicated that profiling LiDAR is an effective tool for regional-scale inventory. In order to quantify the variance of the areal estimates from profiling transect laser data, Nelson et al. (2008) examined three variance estimators, namely Simple Random sampling (SRS), successive difference (SD), and Newton's method (NM), by using PALS data acquired over 56 flight lines spaced 1km apart; they found that the variances of the three estimators are conservative as compared to the systematic standard errors, and they suggested that SD and NM estimators should be used when flight lines are closely spaced (e.g., 2-6 km apart) because autocorrelation between adjacent flight lines is high.

1.3.2 Scanning LIDAR

An expanding body of literature on scanning LiDAR has been witnessed in recent years (McCombs et al., 2003; Riaño et al., 2004; Næsset, 2004; Popescu et al.,

2004; Andersen et al., 2005; Maltamo et al., 2006; Mutlu et al., 2008). The mainstream of the research has been focused on the applications of airborne laser scanners (ALS) for vegetation assessment, with aims of developing novel and effective methods for mapping such forest structural characteristics as biomass, canopy height, basal area, Leaf Area Index, fuel models, canopy closure and timber volume (Popescu, 2002). We herein only present a general review on relevant literature. Typically, the first step in extracting canopy vertical information is to reconstruct terrain topography of bare grounds because canopy heights are relative to grounds and LiDARs only measure the absolute locations of canopy elements in a geographic coordinate system. For this purpose, many algorithms have been developed to generate Digital Terrain Model (DTM) from ALS data by first filtering out ground laser hits and then interpolating them to a regular grid. Filtering schemes often involve adaptive and iterative local minimum filters, and the interpolation often employ triangulation, inverse distance weighting, and spline, kriging and so on (Kraus and Pfeifer, 1998;Young et al., 2000). Once terrain information is available, a Canopy Height Model (CHM0, which represents the height of canopy top surfaces relative to ground, can be created by taking height differences between laser hits from canopy upper surfaces and ground. CHM has become a standard LiDAR-derived product that proves useful in many forestry applications (Popescu et al., 2003; Zhao and Popescu, 2007).

The use of ALS data for forest inventory has been studied at different scales of interest such as tree, plot and stand levels. Individual-tree delineation has attracted many research efforts when ALS data with a relatively high density became available. Popescu et al. (2003), for example, proposed the use of a local maximum filtering with variable circular window for identifying trees on a LiDAR-derived CHM. Based on the simple assumption that a higher tree has a wider crown, the window size varies adaptively according to canopy heights, and it, in practice, can be determined with the help of a crown-height equation that is fitted from ground measurements over the study area. This algorithm, as implemented in a software package called TreeVaw, has received many applications for various forest types. Chen et al. (2006), adopted a morphological

analysis scheme to segment CHM for individual tree isolation, and their algorithm proves effective for delineating irregular crown shapes.

Estimation of forest structural characteristics above tree levels typically relies on establishing regression models to relate forest biophysical variables of interest to sensible LiDAR metrics that need to be carefully selected, and researchers in LiDAR remote sensing for forests have been striving to choose the most appropriate LiDAR-derived metrics and forms of model for a specific application. As described in Naesset (2002), standard practice in building prediction models follows a two-stage procedure whereby regression models, either linear or nonlinear, are first developed and trained based on LiDAR data and coincident ground measurements, and then the trained model, upon validation, can be applied to the rest of data to predict forest structural variables. Most commonly used LiDAR metrics include but are not limited to a variety of height metrics such as mean, maximum, median, quantile-based, and quadratic mean canopy heights as well as different canopy density metrics. Lim et al. (2004), for example, examined the utility of canopy-based quantile height with a linearly-transformed multiplicative model for predicting above-ground biomass, by assuming that canopy height distributions are an implicit function of leaf foliage distribution. Their findings suggest that any LiDAR-based quantile heights have similar prediction performances, provided that the allometry remains the same across forests. Previous studies also demonstrated the effectiveness of LiDAR-derived metrics for estimating other forest structural variables such as LAI, basal area, and timber volume across a wide range of forest conditions.

Airborne and spaceborne large-footprint waveform LiDARs have also been employed to characterize forest structure. These instruments include two airborne sensors, i.e., the Scanning LiDAR Imager of Canopies by Echo Recovery (SLICER) and the Laser Vegetation Imaging Spectrometer (LVIS), and one spaceborne Geosciences Laser Altimeter System (GLAS) onboard the ICESat satellite (Lefsky et al., 1997; Lefsky et al., 1999; Means et al., 1999; Drake et al., 2002; Sun et al., 2008). Lefsky et al. (2002), for example, found that LiDAR-measured biomass estimated from SLICER data

using a regression method can explain 84% variation in the ground measurements made at sites in three biomes; their results suggest that large-footprint waveform LiDAR is capable of measuring canopy structure in moderate to high biomass forests.

1.4 Research Problems

1.4.1 The Need for Affordable and Portable Remote Sensing Systems and for Repeatable Observations

Although scanner lasers have already gone far beyond the proof-of-concept phase and are attracting considerable research efforts, there still exist gaps between the plethora of scientific research and the practical applications, especially for large-scale forest inventory. Although these gaps are shrinking, ALS data are expensive and data loads can be appreciable (Nelson et al., 2004). The consequent post-processing and analysis for information extraction would be equally computationally consuming. On another hand, most analysis software is proprietary; and user friendly freeware or software routines that make these data more accessible are lacking. In addition, due to the sophisticated requirements for hardware systems like Inertial Measurement Units (IMU) or Inertial Navigation Systems (INS), a laser scanner may maintain a price tag in the range of millions of dollars plus or hundreds of thousands of dollars, depending on the level of sophistication. The high costs also exclude the operational foresters from the use of such systems. Moreover, such systems sometimes can be prohibited from being exported outside the homeland because of security policies, and they often are married to particular airborne platforms, or have certain special requirements for the platform configurations. The lack of portability and high costs often make ALS systems impractical for data collection over remote areas such as the Amazon, the circumpolar boreal forests, and the Southeast Asia. Moreover, monitoring changes in forest conditions over time is essential to better understand the functional responses of terrestrial ecosystems to global changes, and this requires periodic, repeated

measurements (Turner et al., 2004). However, the limitations of ALS, e.g., high costs, lack of portability and intensive data handling, render repeated observations difficult.

As such, the advantage of acquiring complete coverage using an airborne scanner is somewhat offset by data load, the needs for complicated post-processing, and the expense of the scanning system and data acquisition. However, a profiling system, to certain extent, can complement all these weaknesses. First, the hardware specification of a profiler is relatively simple, thus making the assemblage cost affordable e.g., a total cost of about US$ 30000 for PALS (Nelson et al., 2003). Because profiling data are collected only along individual flight lines typically several kilometers apart, the data volume is much reduced. Profiling measurements are in the form of 1-D height profile, so the computational demand for data processing and analysis is low and the development of algorithms is relatively easy. The high portability of the PALS allows for conducting forest inventory worldwide, especially over some inhospitable areas. With all these attractive features, it is practical to deploy a PALS-like profiling laser system for missions of repeated data collection. Of particular note is that data from profiling lasers capture canopy height profiles only along the ground track of the aircraft, hence, containing much less information than those of scanning lasers. Thus, it is impractical to use 1-D profiling measurements to estimate forest characteristics at the individual tree level, but this constraint is of little concern if the interest is in inventorying forest resources at relatively large scales, i.e. a level of counties, states, the nation or even the continent.

1.4.2 The Need for Real-time Remote Sensing Platforms

A large number of Earth Observing Systems are currently operating, producing a considerable volume of digital data per day (EOS website). The huge data volume presents a new challenge for remote sensing applications in terms of data processing and information extraction (Lavender and Groom, 1999; Masayuki and Hidefumi, 2002). Traditional techniques of processing and distributing remote sensing data may not be sufficient to meet the needs of many end-users, in particular for some emergent

applications that need timely information. The conventional work flow for a specific remote sensing mission is illustrated in Figure 1.3. Participation of different sectors is needed at different phases of a mission, and protocol and cooperation are often necessitated to facilitate the transition from one phase to another phase. The completion cycle of a mission largely depends on the nature and purpose of the mission, which may ranges from weeks, to months, or even to years.

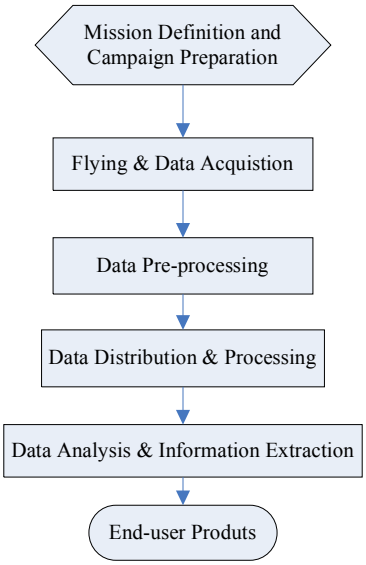


Figure 1.3. A simplified version of typical workflow for remotely sensed data processing

One important aspect concerning the efficient use of remote sensing data is a streamlined flow for data analysis and information extraction, preferably on a real-time basis. The design and implementation of a real-time system highly rely on the automation of conventional procedures. Different levels of automation exist for a task, depending on how much hardware and software facility is available to fulfill the task. To implement the workflow for a specific task, a level of automation usually needs to be defined with respect to the objective of the task. In practice, the development of such real-time systems is far from an easy endeavor because this requires high levels of automatic data processing and robust information retrieval algorithms so that human

intervention and manual inputs are minimized as much as possible (Derrien et al., 1993; Pearson et al.,1994; Suter and Nuesch, 1995; Chang et al.,2001). Although such systems are desired in many sectors, the related researches and investigations are still at an initial stage with a lot of obstacles to be overcome.

1.4.3 Motivation and Objectives

At present, real-time processing and information extraction from remotely sensed data represents a novel tendency in the development of new remote sensing systems. In the context of forestry remote sensing, a platform capable of real-time delivery of forest information is of practical significance, and will have attractive features. It will enormously enhance the speed of information delivery, and save a considerable amount of labor needed for post-processing and analysis. The quick access to the forestry information not only facilitates the process of decision-making but also provides capabilities of meeting urgent needs, e.g., in cases of natural or human-induced hazards such as forest fires (Li et al., 2000).

Given the interest in real-time systems and the attractive features of profiling laser systems for forest inventory, the primary objective of this research is to develop a near-real time profiling LiDAR-based forest inventory and mapping tool that delivers end-user products within hours of the final flight of the mission. In order to implement such a system, the specific sub-objectives involved are as follows:

(1) To build a paradigm for processing profiling LiDAR data on-the-fly, as a step toward real-time estimates of biophysical parameters.

(2) To develop algorithms for automatically analyzing and segmenting LiDAR height profiles acquired by the system

(3) To develop accurate models, whether statistical-based or not, for estimating forest biophysical parameters, such as canopy height, basal area, merchantable volume, biomass, above-ground carbon etc., from the data of canopy height profile; and to explore the possibilities of extracting other information from the profiling data.

(4) To investigate the accuracy of the profiler-derived estimates, and compare

them with the scanner-derived products and field measurements; therefore, part of efforts will be devoted to investigating new methods of extracting forest characteristics from scanning LiDAR data.

(5) To investigate the potential of the fusion of profiling data with remotely sensed data from other platforms, whether airborne-based or satellite-based. One possibility is to use profiler-derived product as a quickly accessible source to calibrate the satellite-based products.

(6) To implement all the developed algorithms as computer routines, and integrate them into the computer packages to be consistent with hardware components of the profiling system.

1.5 Dissertation Organization

The successful implementation of the proposed on-the-fly airborne laser profiler relies on streamlining and integrating the conventional workflow by automating data analysis and information extraction. As such, the dissertation is organized in such a way that each chapter deals with a relatively independent issue that is associated with different phases of the workflow of a remote sensing application system. Specific topics for each individual chapter are summarized as follows:

(1) Chapter II presents a framework of automating PALS as an on-the-fly reconnaissance tool for forest inventory over large areas, and also details the ground-finding and segmenting algorithms for automatic processing PALS 1-D height profile data.

(2) Chapter III attempts to develop variance estimators to quantify the uncertainty in LiDAR-based estimates of forest characteristics; a case study in estimating canopy percentage cover is examined.

(3) Chapter IV aims to develop prediction models for above-ground biomass. The proposed models feature scale-invariance and they can be applied to both profiling and scanning LiDAR measurements.

(4) Chapter V describes models to estimate Leaf Area Index from both profiling and scanning LiDAR data. The potential of integrating satellite products such as MODIS with PALS is also examined.

(5) Chapter VI describes the software development that implements the data processing algorithms and prediction models of the previous chapters.

(6) Chapter VII, the concluding remarks, summarizes the research of the dissertation as well as points out the possible directions for future studies.

CHAPTER II
AN ON-THE-FLY AIRBORNE PROFILING LIDAR SYSTEM: PARADIGM AND ALGORITHM DEVELOPMENT

2.1 Introduction

Advances in computational power and memory open up new opportunities to build an integrated data acquisition system with near real-time derivation and delivery of end-user products (Sandau et al., 2000). Near real-time or real-time remote sensing systems can provide unique capabilities for a variety of natural resources-related applications, especially those that need timely responses, such as monitoring and controlling natural disasters and man-induced forest fires (Honda and Nagai, 2002**;**Ramsey et al., 2004). On the other hand, an enormous amount of remotely sensed data becomes available as many on-orbit platforms are routinely operated and many airborne flight missions are frequently scheduled. The huge data volume often precludes manual-based processing, and in such a sense, hardware and software systems dedicated to automatic processing and analysis of remote sensing data are in need (Hsu et al., 2001; Shutler et al., 2005; Tehranian et al., 2006).

Although most real-time systems require on-board data processing modules, no accepted architecture and techniques exist for such systems (Konare et al., 2003; Plaza, 2008). The design and implementation of a real-time remote sensing system are often mission- or case- specific, largely depending on the availability of hardware/software and budgets as well as the purposes of applications. For example, Toth and Grejner-Brzezinska (2002) proposed a prototype mobile mapping system for automatically identifying road centerlines and extracting image features. Their system, developed as an integrated GPS/INS/CCD system, features automatic and real-time image processing so that no human interaction is needed. Lienert et al. (1999) developed a software system for real-time analysis and visualization of scanning LiDAR data for deriving information on marine aerosol, and they concluded that the software can be adapted to other LiDAR

systems such as Differential absorption LiDAR and oceanic LiDAR. In addition, a journal called "Journal of real-time image processing" was recently started specifically for dealing with designing, implementing, and utilizing of real-time image/video processing systems.

Recently, LiDARs, most notably airborne laser scanners, have brought a breakthrough in forestry remote sensing (Lefsky et al., 2003). However, due to the sophisticated hardware, high costs, huge data volume, and complicated processing procedures, it seems quite impractical to develop an on-the-fly scanner system for real-time forest inventory, especially over large forested areas (Nelson et al., 2004). In contrast, a laser profiler such as the portable airborne profiling LiDAR system designed by Nelson et al. (2003a) -- featuring technical simplicity, transportability, and manageable data volume -- somehow offsets scanning systems' disadvantages, and thus can serve as an archetype to develop an on-the-fly LiDAR system. As pointed out by Nelson et al. (2004), it is expected that airborne profiling systems will be automated so that they are flown to produce estimates of timber, biomass, and carbon upon landing of the aircraft, and that an automated scanning system follows soon afterwards. To this end, this research aims to develop such an automated profiling LiDAR system that processes data in-flight and estimates inventory information upon landing. This chapter is focused on presenting a paradigm framework as well as relevant technical details for the on-the-fly profiling LiDAR system.

2.2 Paradigm of the on-the-fly Airborne Laser Profiler

This research is based on the earlier development of profiling laser systems. Several previously developed laser profilers prove effective and useful in forest inventory (Ritchie, 1995; Tickle et al., 1998; Nelson et al., 2003a; Millette and Hayward, 2008). Therefore, no attempt is made in this research to assemble a profiler from scratch; instead, the PALS system developed by Nelson et al. (2003a) is employed as a prototype to develop the on-the-fly system because at present, PALS is one of the most widely

used profilers that has been flown aboard various aircrafts worldwide over areas such as the sates of Texas, Delaware and New Jersey of USA, Japan, and Norway. Previous studies also demonstrated that PALS is an effective and affordable tool for large-scale resources inventory (Nelson et al., 2003b; Nelson et al., 2005; Skowronski et al., 2007). In this regard, PALS provides a good prototype system upon which to build for an automated laser profiler.

For completeness, a brief description of PALS is presented below. PALS was assembled from off-the-shelf, commercially available components at an approximate cost of $30,000 (Nelson et al., 2003a). Its major components are as follows (Figure 2.1):

(1) A near-infrared laser transmitter/receiver to measure first-return ranges and magnitudes from laser to targets;

(2) A Global Position System (GPS) receiver;

(3) A charge-couple Device video camera with GPS video titling so that time/location synchronization information is available;

(4) A laptop computer to record data streams from the dGPS and laser receiver;

(5) And a commercial software implemented in LabView to control, monitor and record the data streams.

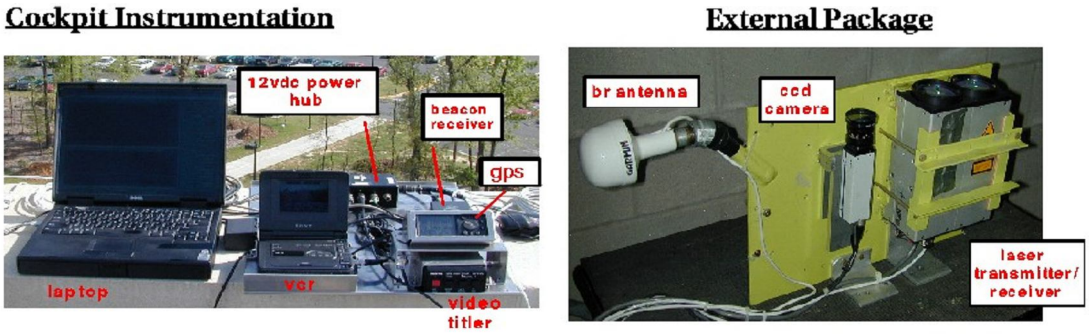


Figure 2.1. The major off-the-shell components of the portable airborne laser system (PALS) (Photos courtesy of Dr. Ross Nelson of NASA).

PALS was designed to fly over an altitude up to 300 m above ground. The

transmitter, in its original configuration, has a pulse rate of 2000 hz, a 0.2-mr divergence and 10-cm optics, thus resulting in a 0.4-m spot at a flight height of 150 m. The system as configured currently can record sequential first, last ranging returns at repetition rates ranging from near zero up to 2000 hz. Typically, the recording rate is set to provide a continuous trace of spatially adjacent pulses based on the nominal speed of the host aircraft. For practical applications, any component of the PALS system can be replaced or upgraded to meet specialized needs, depending on the availability of budgets and the purposes of applications. For example, recently Nelson et al. (2004) reported the upgrade of the PALS dGPS system to enable the Wide Area Augmentation System (WAAS), and this upgrade allows for an aircraft positional accuracy of 1-2 m horizontal and 2-3 m vertical within the continental USA. However, due to electronic lags inherent in the data acquisition system and unknown roll/pitch errors, the locational accuracy of any given pulse will be typically on the order of 10 -15 m.

The scheme to automate PALS for on-the-fly data processing is technically straightforward, with the aim being to perform the data processing and analysis tasks in-flight that originally were done post-flight. A schematic flowchart for the on-the-fly system is presented in Figure 2.2. As noted before, a key point in designing and implementing an on-the-fly real-time system is to develop on-board data processing software that implements computer procedures or algorithms used for analyzing and processing the data simultaneously with the flying. More importantly, for the software to work smoothly with little or no human interaction, the on-board processing algorithms should be robust enough to minimize the human intervention for correcting errors. To this end, two major PALS data-processing algorithms, one for finding ground hits and another for segmenting canopy height profiles, are integrated in a software package for on-board data processing. First, the ground finding algorithm attempts to identify those laser hits that are reflected back from ground, and then, the identified grounds hits can be interpolated to generate a 1-D terrain topography. With the resulting 1-D topography, a canopy height profile can be derived by taking the difference between laser range and the corresponding terrain elevation. Next, the segmentation algorithm is used to stratify

the canopy height profile into strata with each of them representing a different type of land cover, e.g., forest, open ground, and buildings.

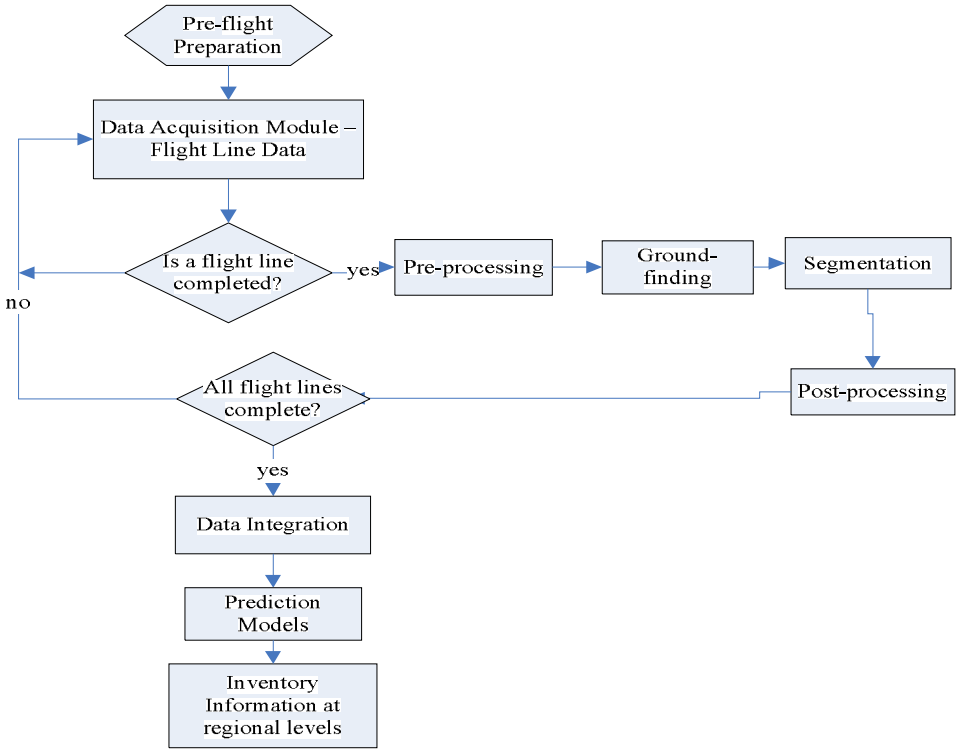


Figure 2.2. A framework of the proposed on-the-fly PALS profiling LiDAR system.

Prediction models, i.e., regression models that use LiDAR-derived metrics as predictors, are also required to convert the LiDAR measurements into forest characteristics of interest (Næsset, 2002). Often enough, these models are developed based on a training dataset that includes LiDAR-derived metrics and coincident ground-measured values of forest structural variables. If ground data are not measured, it seems not practical to directly train a prediction model. In such a case, it may be helpful to resort to either models that are previously developed for the same area or those not for the same area but under similar forest conditions. In case that no other models can be borrowed at all, a compromise is to develop a computer simulator to generate LiDAR measurements with known forest characteristics and then to use the simulated data to

build and train models, as demonstrated in Nelson et al. (1997). In fact, most research efforts in LiDAR remote sensing for forestry application have been dedicated to looking for the most appropriate LiDAR-metrics and forms of model.

PALS cannot provide a direct areal estimate of forest inventory information, primarily due to the incomplete observations that are collected only along transects under the flight lines as shown in Figure 2.3 (Nelson et al., 2005; Zhao et al., 2008). As such, for regional assessment of forest resources using PALS, standard practice is to rely on Line Intercept Sampling procedure to extend the transect-level estimates of forest characteristics to the whole extent of the study area (Nelson et al., 2004). Obviously, such areal estimates are subject to uncertainties due to its sampling nature, and a measure needs to be formulated to quantify the variance of LIS-based areal estimates (Nelson et al., 2008). In practice, these variances are essential to guarantee the informative use of the PALS-derived estimates.

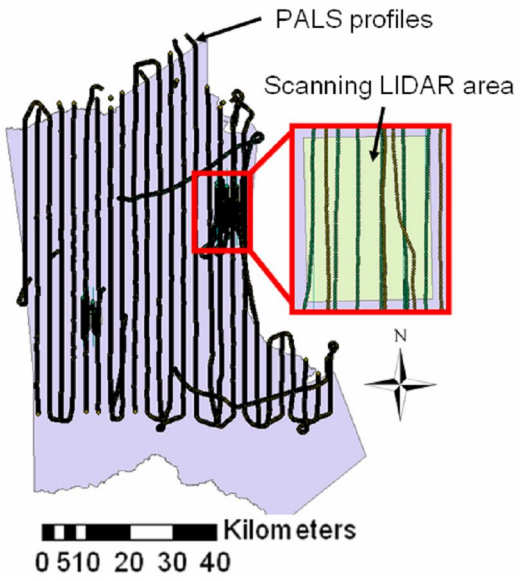


Figure 2.3. The PALS flight line pattern for a mission aiming to assess regional resources in East Texas.

2.3 Data Processing Algorithms

2.3.1 Ground-finding Algorithm

For real-time or near real-time processing of PALS profiling laser measurements, automatic procedures are entailed to process and analyze PALS ranging measurement, as shown in Figure 2.4. In essence, the PALS measurements along a transect (flight line) can be viewed as a 1-D signal where the ranging is plotted against the pulse numbers. First of all, a ground curve, representing the terrain topography along the flight line, needs to be derived in order to compute the heights of sensed objects, e.g., trees, relative to the ground (Figure 2.4). Hence, that subset of near-ground laser hits that are returned back from the actual ground need to be identified. For both profiling and scanning LiDAR data, most algorithms for classifying ground hits are based on local minimum filtering that may often be refined with some adaptive strategies.

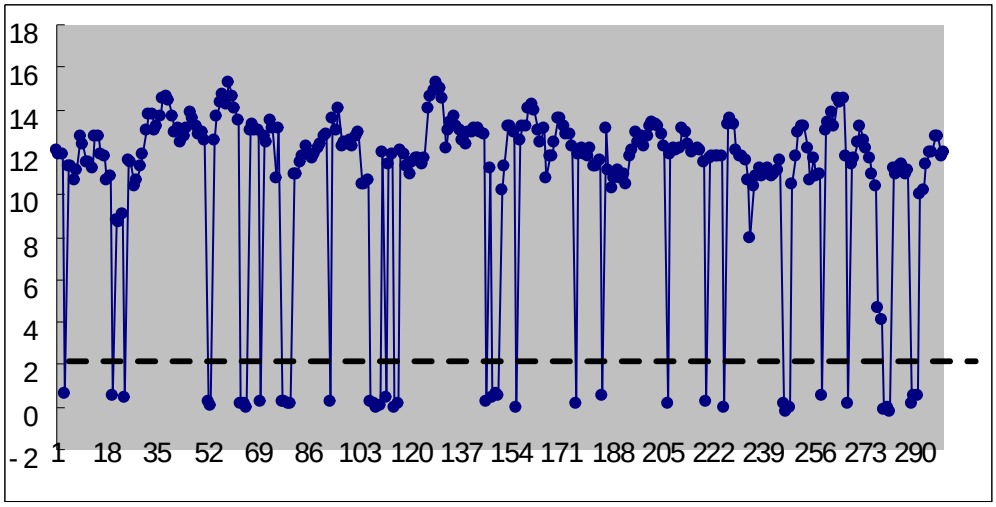


Figure 2.4. A profile of ranging measurements measured by PALS where all the laser hits below the vertical dashed line are supposed to reflect from the ground.

The basic idea behind local minimum filters for ground-finding is that, given that some laser pulses penetrate the canopy to reach the ground, the lowest hits within a neighborhood should be a ground hit. Of note is that the penetration rate of lasers depends on both the laser characteristics, e.g., wavelength, power, and footprint size, and the canopy conditions. A very dense canopy can effectively attenuate all the energy of lasers, so ground cannot be sensed. Therefore, caution should be exercised in designing a local-minimum algorithm.

In the original implementation of PALS ground-finding procedure, Nelson et al. (2003b) used a non-overlapping moving window to locate local minima as ground hits, with the window size defined by analyst. Then a spline interpolator is employed to generate the ground curve. Overall, their algorithm, being intuitively simple, performs satisfactorily except under certain pathological situations, e.g., step functions-cliff, rapid sharp topographic changes, very dense canopies with little or no ground peneration. But there is no correcting mechanism in the algorithm to handle these pathologies. Also, the spline interpolator sometimes causes overshooting or undershooting due to the smoothness constraints of the interpolator. Therefore, the ad hoc post-processing supervised by users is usually needed to correct some conspicuous errors in the ground curve. Such user intervention makes on-board data processing impractical. To improve this algorithm as well as minimize user interaction, some heuristics rules are introduced to make the ground-finding algorithm more robust. The improved algorithm is an iterative procedure as detailed below.

(1) Local Minimum Filtering: An overlapping window, instead of a non-overlapping one, is used to search local minima as potential ground hits. The window size is defined by users. To obtain as many ground hits as possible, a maximum overlap between consecutive windows are assumed; that is, the window is moved pulse by pulse. Obviously, the use of an overlapping window often causes the same ground hit to be identified multiple times by several consecutive windows; in such a case, the ground hit should be recorded only once.

(2) Spline Fitting: With the ground hits picked up, a spline is fitted to generate a ground curve.

(3) Correction Scheme – Adding new ground hits: All non-ground hits are compared against the ground curve generated in Step 2. If a laser hit lies below the ground curve by a height threshold, say, 0.5 m, the laser hit will be labeled as a ground hit. With the expanded set of ground hits including the newly added ground hits and the original ones, iterate Step 2 and Step 3 until no ground hit is added.

(4) Correction Scheme – Eliminating mis-labeled ground hits: Split the ground hits obtained after Step 3 into five subsets by taking every 5^{th} ground hit. Combine any four subsets to generate an interpolated ground curve, and compare the resulting curve with the remaining subset of ground hits. If a ground hit in the remaining subset is above the ground curve by a height, e.g., 0.5m, this hit is considered as a misclassified ground hit and is removed from the set of ground hits. After evaluating all five combinations of the five subsets, an updated set of ground hit is created. Then, with this new set, repeat Steps 2, 3 and 5 until no ground hits is added and eliminated, or a pre-set maximum number of iteration is reached.

Once the ground curve is available from the above procedure, heights of objects relative to the ground can be directly obtained by subtracting the ground topography from the original height profile. The resulting height profile is called canopy height profile (CHP) (Figure 2.5). Be cautious that the same term "CHP" was also used in large-footprint LiDAR literature to represent corrected LiDAR waveforms. It, therefore, may be necessary to unify and clarify terminologies to avoid possible confusion in the community of LiDAR remote sensing.

2.3.2 Segmentation of CHP

Canopy height profiles need to be segmented with respect to land cover types. Since flight lines are typically ten's or hundreds of kilometers long, not all profile measurements come from forest. Figure 2.5, for example, shows that a CHP may be split into four segments, each of which corresponds to one land cover class, i.e., open

ground, man-made building (residence area), mixture wood and pine. For forest inventory, at least two strata should be differentiated, i.e., forest, non-forest. Although the further segmentation of non-forest portions of CHP into more detailed strata is desirable, it remains difficult to rely only on the 1-D transact height measurements to extract strata information, particularly in an automatic manner. A remedy for this may be to refer to some auxiliary data such as optical images, if available. In fact, the CCD video camera of PALS is employed to record overhead images of surfaces that could later be used to determine the strata of CHP by matching CHP with the videos; and the matching is made possible because videos have been synchronized with the ranging measurements of CHP. However, this matching process is done visually and manually by analysts, and usually very labor-extensive and time-consuming. Because the automatic analysis of video contents is rather difficult, the algorithm proposed in this research for segmenting CHP will not take the video into account.

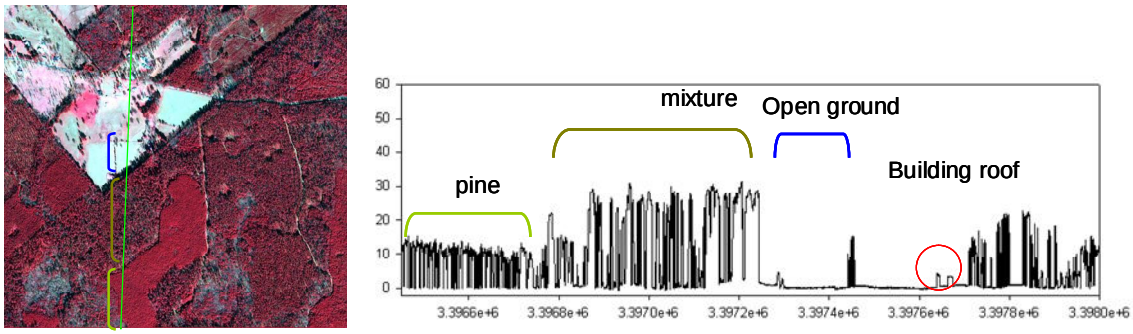


Figure 2.5. Segmentation of a LiDAR-measured canopy height profile with respect to land cover types: the left is a false-colored Quickbird image overlaid with a profiling LiDAR flight line (the green line), and the right is the height profile corresponding to the transact denoted by the green line.

To develop an algorithm for segmenting CHP, this research resorts to the stationary wavelet transform to perform segmentation on wavelet domain. For ease of reference, the basic theory of stationary wavelet transform is described in brief below.

Wavelet analysis is a powerful tool commonly used to analyze 1-D digital signals. With a forward wavelet transform, a signal can be decomposed into a series of coefficients in the scale-translation domain. By appropriately enhancing or reducing magnitudes of certain coefficients in the new representation of the signal, the reconstructed signal through the inverse wavelet transform applied on the modified coefficients can reveal certain desirable features or suppress undesirable features. Specifically, for digital signal processing, discrete wavelet transform (DWT), i.e., a discretized version of wavelet transformation, is probably the most extensively used algorithm. DWT can be mathematically expressed as,

$$f(t) = \sum_{j \in Z} \sum_{k \in Z} d_{j,k} \psi_{k,k}(t) \tag{2.1}$$

where $f(t) = \in L^2(R)$ is a 1-D square-integralable function; $d_{j,k}$'s are the DWT coefficients, and $\psi_{j,k} = m^{j/2} \psi(m^j t - k)$ are the basis functions with $\psi(t)$ being the mother wavelet function and $m = 2$. Because $\psi_{j,k}$ are the scaled and translated versions of $\psi(t)$, DWT provides a multi-scale representation of $f(t)$. In practice, DWT, however, suffers from a prominent drawback of being translation-invariant, i.e.,

$$T \cdot DWT[f(t)] \neq DWT[T \cdot f(t)] \tag{2.2}$$

where T is a translation operator and DWT represents the DWT transformation operator. Eq. 2.2 means that when performing DWT, the transformation results depend on the choice of the origin. To compensate this drawback, Stationary wavelet transform (SWT) is proposed to achieve the translation-invariance. SWT is quite similar to DWT except that SWT does not decimate when applying the high and low filters on signals at one level to produce two new sequences at the next level. Accordingly, the length of coefficients at a level is equal to that at the previous level, and the sequence length of decomposed coefficients at each level is equal to the length of the original signal. However, to ensure performing the filters appropriately at each level, the filters in SWT need to be padded with extra zeroes between each pair of adjacent elements. For example, suppose at the level of j the filter elements are:

$$\cdots, h_0, h_1, h_2, h_3, \cdots$$

Then, the filter elements at the next level of $j - 1$ should be,

$$\cdots, 0, h_0, 0, h_1, 0, h_2, 0, h_3, \cdots$$

At the same time, a downsampling needs to follow before performing the filter at next level. In principle, SWT has a close relationship with the e-decimated DWT , but SWT contains the coefficients of e-decimated DWT for any choice of e. Moreover, it becomes clear that SWT is a redundant transformation. This very redundancy, however, facilitates the detection of salient features in signals. SWT finds important applications in signal processing and analysis, e.g., wavelet shrinkage. More details on SWT can be found in Nason and Silverman (1995) or Coifman and Donoho (1995).

In a LiDAR-measured CHP, the height variations over different land cover types may show distinct patterns. For example, heights over a forest segment have more variations than those over the other land cover while the height variation of open ground is supposed to be zero. But, the actual variation in height of open ground often is nonzero due to the random measurement errors as well as the inaccuracy in deriving the ground topography. Man-made buildings with linear or flat roofs are also supposed to show little variation. Therefore, the pattern in height variation provides hints for segmenting CHP, and one possible way to capture these variations (detailed structure of the height profile) is to examine the detailed component of the wavelet transformation. To guarantee that the coefficients in wavelet domain are mapped to the positions of original data one by one, SWT is a preferred choice because it can preserve the signal length during the transform.

The procedure to segment CHPs is technically straightforward as illustrated in Figure 2.6; and the following steps are involved:

(1) Decompose a CHP by SWT to obtain the 1st-level detailed component that captures the high-frequency variation in the CHP.

(2) Take the absolute value of 1st-level detailed component to get positive wavelet coefficients, and then apply a mean filter to the positive detailed component to obtain a smoothed version of the detailed signal.

(2) Threshold the smoothed detailed signal using a cut-off value.

The resultant binary signal from thresholding is the segmented profile. All the 1's in the binary signal are considered as forest. The zero-valued points need to be further classified by referring to the associated height values: those 0's in the thresholded profile that have a height above 3m will be considered as roofs, and otherwise, the 0's are considered as open ground.

As to the aforementioned procedure, several points need to be clarified. First of all, in Step 1, the form of wavelet must be specified to perform SWT, and the Daubechies wavelets usually are good choices. In Step 2, the purpose of mean filtering before thresholding is to ensure that a more likely continuous segment is obtained. In Step 2 and 3, several parameters involved need to be specified empirically by users, including the window size of mean filtering, the cut-off value in thresholding, and the height threshold in differentiating between ground and roof.

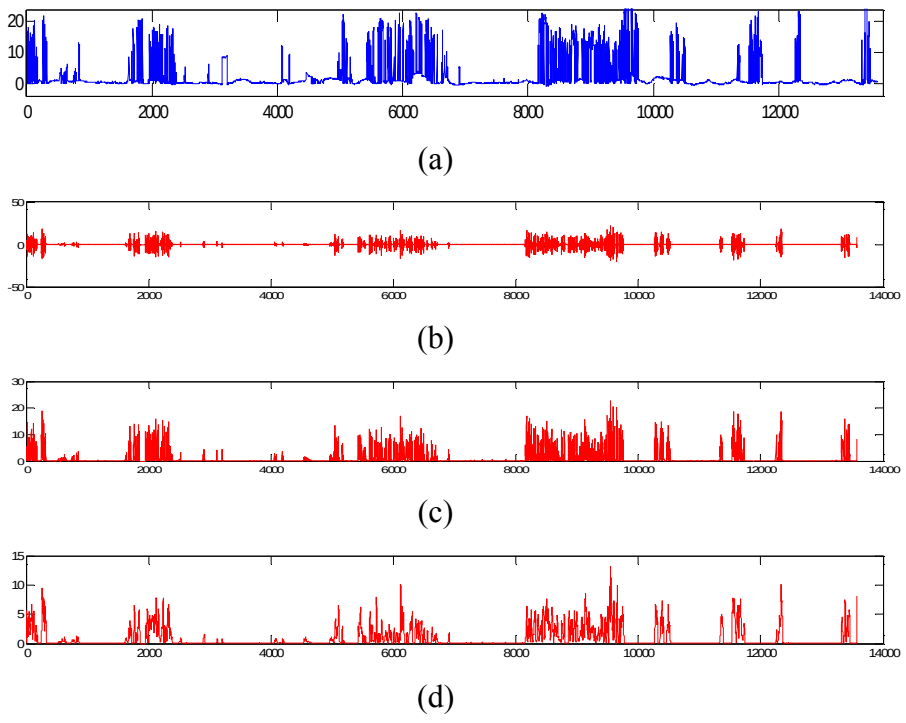
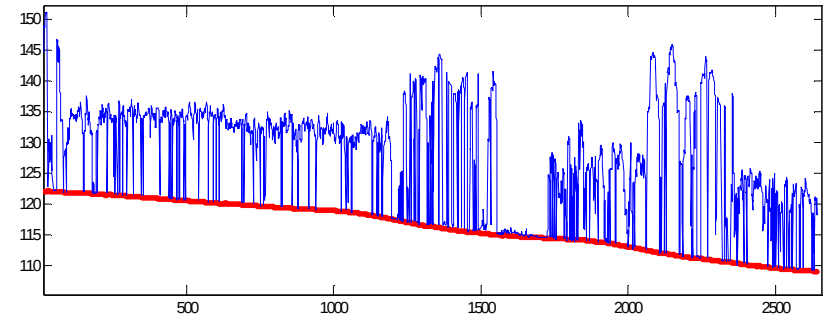


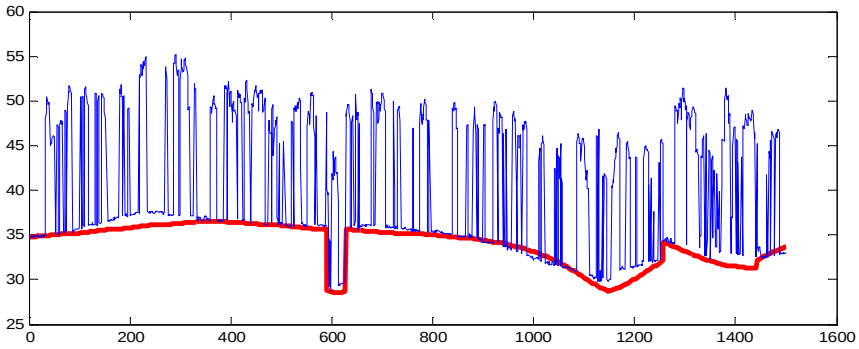
Figure 2.6. The SWT decomposition of height profiles and the subsequent analysis for segmentation: (a)The original height profile, (b)the 1st-level detailed component by SWT, (c) the absolute value of (b), and (d) the signal obtained by applying a mean filter to (c).

2.4 Results and Discussion

The ground-finding algorithm was applied to the PALS data collected over an intensive study area of East Texas, i.e., the highlighted rectangle area in Figure 2.3, where coincident scanning LiDAR observations collected by a Leico-geosystems ALS are also available. In the algorithm, a window size of 100 m was used for local minimum filtering, and a maximum of 10 iterations was set for the iterative searching. Two typical examples of the resulting ground curves are shown in Figure 2.7. Over flat and monotonically decreasing or increasing topographies, the ground curve can be accurately recovered, as depicted in Figure 2.7a.. However, the searching for ground hits is a little problematic over arched terrains (e.g., the leftmost portion in Figure 2.7b). Such failure is an inherent nature of local minimum filtering due to its incapability of locating minima over a concave topography that extends longer than the window size; the situation can be mitigated if a smaller window is used. But too small a filter window cannot guarantee the presence of a ground hit within the filter. Therefore, in practice, a balance should be taken for choosing an appropriate filter window size. In particularly, if forest conditions are known so that the penetration characteristic is understood, such prior knowledge can guide the selection of a filter as small as is allowed for the forests.



(a)



(b)

Figure 2.7. PALS-measured canopy height profiles as well as the ground curves derived by the ground-finding algorithm, for two different terrain topographies.

The PALS canopy heights over the intensive study area were derived after the ground curves became available. Then, these PALS-derived canopy height measurements were compared to those obtained by the scanner, (1) to directly evaluate the accuracy of PALS measurements using scanner observations as references, and (2) to indirectly examine the effectiveness of the ground-finding algorithms since the PALS canopy heights were obtained by using the reconstructed ground curve. A typical result is presented in Figure 2.8, with a mean canopy height of 7.22 m for PALS as compared

to 8.82 m for the scanner; a two-tailed p-test suggests that the difference between the two values are insignificant (p-value << 0.05). The difference is suspected to mainly result from the misregistration between the PALS with the scanner.

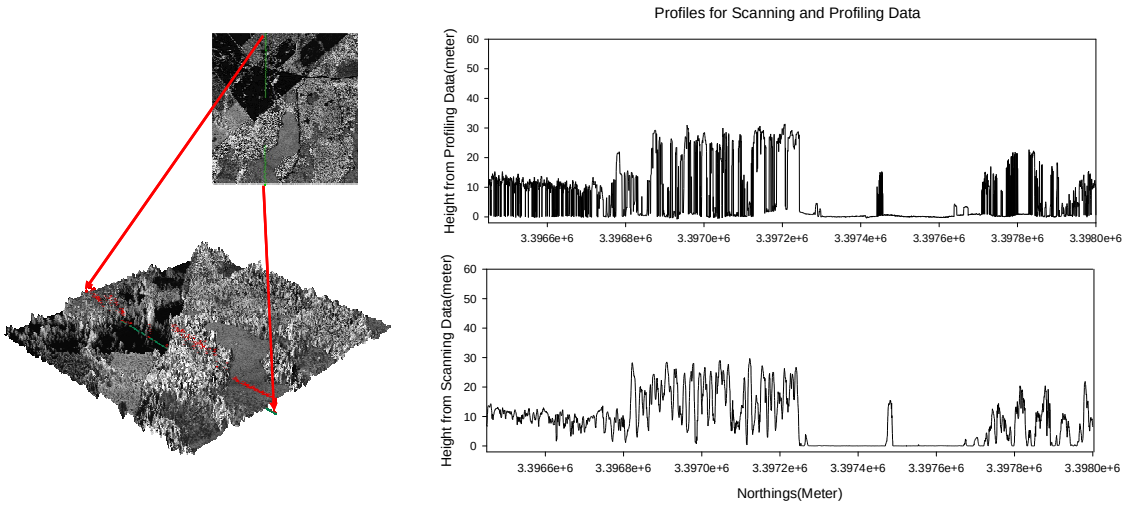


Figure 2.8. A visual comparison of a PALS height profile with the associated scanner-derived profile: a scanner-derived CHM, scattered with PALS laser hits, over a subset of the intensive study area (the left), and the PALS canopy profile (upper right) as well as the scanner-derived profile (lower right)

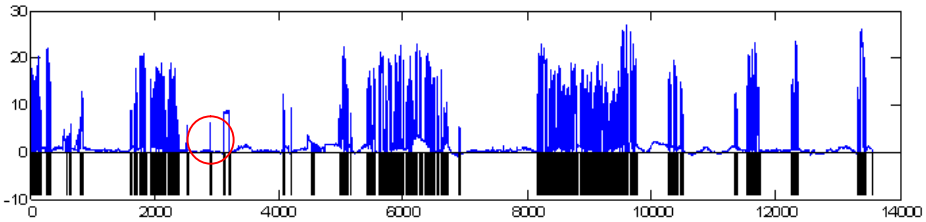


Figure 2.9. The segmentation of a profile where the downward bars indicate the forest segments (excluding those back bar at the building edges).

To segment the height profile signal in Figure 2.6, the SWT with the Daub8 wavelet is applied. The window size of mean filter is chosen to be 25.0 m; the cut-off value for thresholding is set to 0.7 m (0.35 is the nominal error of the laser measurement). The obtained segment result is shown in Figure 2.9 where the downward bars indicate the forest segments (excluding the building). A visual examination on the segmentation result suggests the result is reasonable. Notice here that the forestry segments are not classified further as mixture, pine, or hardwood; and this should be taken into account in the future work. One possible way is to examine the differences of the autocorrelation structures for each classes. In addition, more sophisticated schemes of thresholding will also be exploited. As a further examination of the segmentation algorithm, the ratio of forest strata segmented from the PALS profiles to the total length of the profiles was used as an estimate of forest percentage cover. The estimated value, 79.72%, was comparable to a reference value of 76.14% that was obtained from the classified Quickbird image.

Recently, Maeda et al. (2008) developed a computer program to automatically overlay PALS-measured profiles with coincident videos simultaneously for on-screen visualization based on which to stratify the profiles. The procedure significantly reduces the amount of time invested for segmentation as compared to a conventional approach that uses a manually-operated video player; however, it still remains labor-intensive because analysts are involved to label the profiles according to the visual contents. On the other hand, it may be helpful to incorporate a GIS land map to directly stratify PALS profiles. But a concern may arise that the positional inaccuracy of PALS measurements causes the profiles to be segmented incorrectly due to the misregistration error, as noted in Nelson et al. (2004) where they found the PALS segments falling on the water class of the GIS map sometime do not have a zero height. A more viable alternative is to employ an optical multi-spectral sensor to capture high-resolution mulispectral images concurrently with the laser measurements in that the information for the two types of measurements complements each other. For example, Millette and Hayward (2008) built

such a system with both laser profiler altimetry and multispectral imaging, their system proves useful in mapping detailed forest characteristics. However, an integrated system like this still maintains a high price tag, thus, somehow precluding its use for regional forest inventory over large-scale areas within a reasonable budget.

2.5 Conclusions

Using PALS as a prototype, a framework for designing an on-the-fly profiling LiDAR system was proposed. The PALS system is a portable airborne laser profiler that has been widely used for assessing regional forest resources, and the proposed on-the-fly LiDAR system not only carries on the full capability of PALS, but more importantly extends PALS to produce timely forest inventory information upon landing. Aimed for near real-time inventory, the proposed system features on-board data processing. Two key data processing algorithms as to profiling measurements, i.e., ground-finding, and segmentation, are described in detail. The evaluation of the two algorithms using PALS measurements over a study area where coincident scanning LiDAR data were also available shows that the algorithms generally perform well. Although more complicated algorithms can be investigated to further improve the robustness of the algorithms in later studies, the current versions are sufficient for on-board data processing with little or no human interaction as demonstrated by the result comparisons. The proposed on-the-fly profiling system is designed to be a regional assessment tool, one that can be used to repeatedly, remotely measure hundreds or thousands of square kilometers with little/no analyst interaction or interpretation. It is also envisioned that an airborne laser system integrated with an airborne multi-spectral optical sensor, though with a high price tag, brings complementary information valuable to improve the understanding of forest characteristics; and the on-the-fly system proposed in this research could serve a paradigm for automating the integrated profiling laser systems or even scanning LiDAR systems.

CHAPTER III
VARIANCE ESTIMATORS OF PERCENTAGE COVER USING AIRBORNE PROFILING LASER WITH LINE-INTERCEPT SAMPLING

3.1 Overview

The line-intercept sampling (LIS) method has found important applications in such areas as forest and wildlife, ecological and biological sciences, and crop and agriculture fields. LIS is a sampling technique to make observations along line transects in order to make inferences of area properties. The placement of transects can be chosen in many different manners, i.e., randomly or systematically. The motivation of this study is to use LIS to infer regional information of forestry biophysical parameters based on the linear transect measurements of a profiling LiDAR system. However, there is no optimum method to properly derive a reasonable measure to the uncertainty of LIS estimates. As such, the study first developed a theoretical framework to describe the LIS estimation in two settings, one with fixed landscape configuration, and another with random configuration. The subsequent simulation of transect observations is realized for two categorical maps: the artificial one simulated by SIMMAP, and the real one classified from Landsat ETM+ multispectral imagery. The simulated samples were used to test four estimators. The methodology employed in this study provides a good starting point for practically implementing the quantification of variance estimates with LIS.

3.2 Introduction

The line-intercept method is a quite effective sampling technique, and it has found extensive applications in natural resource-related disciplines, such as the characterization of landscape patterns and the inventory of forestry tracts. Instead of intensive sampling over the whole study area, Line-Intercept Sampling (LIS) measure samples only along linear transects. In practice, the placement of transects over study

areas can be done either randomly or systematically with fixed or varying length for each transect, and sometimes be done in a prescribed configuration. Basically, LIS is used to observe a region consisting of randomly distributed particles or items by only measuring the particles intercepted by transects. Particles under investigation can be geometrically arbitrary-shaped. Based on the settings of applications, the particles may represent different entities such as shrubs, tree crowns, dens, lakes, plants, roads and signs, and patches in a landscape. So far, many estimators have been developed to extend line transect observations about particles into the inference of area properties, e.g., estimating the percentage cover of certain patch type based on the fraction of the transects intercepted by the given type of patch (Kaiser, 1983; Butler and McDonald, 1983).

Historically, LIS technique can be dated back roughly to as early as 1868 when Crofton (1868) discussed the local probability of straight lines drawn randomly in a plane. Canfield (1941) first introduced this technique to sample range vegetation; he employed the random placement of transects and used the proportion of the sampled transect intercepted by the vegetation as an estimate of vegetation percentage cover. Bauer (1943) compared the relative efficiency of the transect and quadrate methods of sampling vegetation based on a series of simulated plant communities of known composition, and he showed that the transect sampling gave more accurate results than quadrate methods in the cases he investigated. In 1953, McIntyre furthered the use of the LIS technique for estimating plant density. Later, Warren and Olsen (1964) proposed a LIS-based estimator for the volume of logging waste and documented the first use of LIS in forestry applications, and instead of calling it line-intercept sampling, they coined the term line-intersect sampling for this technique. The unbiasedness of Warren and Olsen's estimator was proved true under certain assumptions by Van Wagner (1961). Later, Kaiser (1983) presented some theoretical work by introducing two generic estimators for LIS, and his theory tends to unify all the LIS results up to that time. He employed a geometrical probability approach to prove the unbiasedness of estimators he proposed, and provided examples in which his estimators degenerated to some

commonly known estimators. Kaiser's development (1983) proved to be one of the most appreciated works in the literature of LIS. More recently, extensive applications of LIS have been seen in such areas as forestry, wildlife, ecological and biological sciences, and crops and agriculture sciences (Keller et al., 2004; Keane et al., 2005; Nelson et al., 2005). Recent progress of LIS sampling theory can also be witnessed in the work of Affleck et al. (2005), among others.

In this study, we are interested in developing an on-the-fly portable airborne laser profiling system for forest inventory based on an existing system called Portable Airborne Laser System (PALS) assembled by Nelson et al. (2003a). The proposed system is a profiling laser altimeter or LiDAR (Light Detection And Ranging). A laser profiler usually emits laser pulses at near-nadir direction (without scanning) and collects the returned pulses. The round-trip travel time of the laser pulse can be converted to ranging measurements according to the speed of light. As shown in Figure 3.1, a profiling LiDAR can measure the forest canopy height profile along a transect that is defined by the flight line (Nelson, 1984). The schematic in Figure 3.1a suggests that two profiles are involved to characterize the stand structure: the top one depicting the profile of top canopy surface, and the bottom one depicting the terrain topography. Previous studies found that forest stand characteristics can be derived from this profiling measurement (Maclean and Krabill, 1986). Transects flown over a study site in Figure 3.2 illustrate the basic idea about how a profiler collects samples and what the placement of transects looks like. Unlike most current commercially available scanning LiDAR systems capable of continuously sampling the study area (Lefsky et al., 2002; Popescu et al., 2002; Popescu et al., 2003), the proposed profiler can make measurements only along selected, predefined linear or near-linear flight transects.

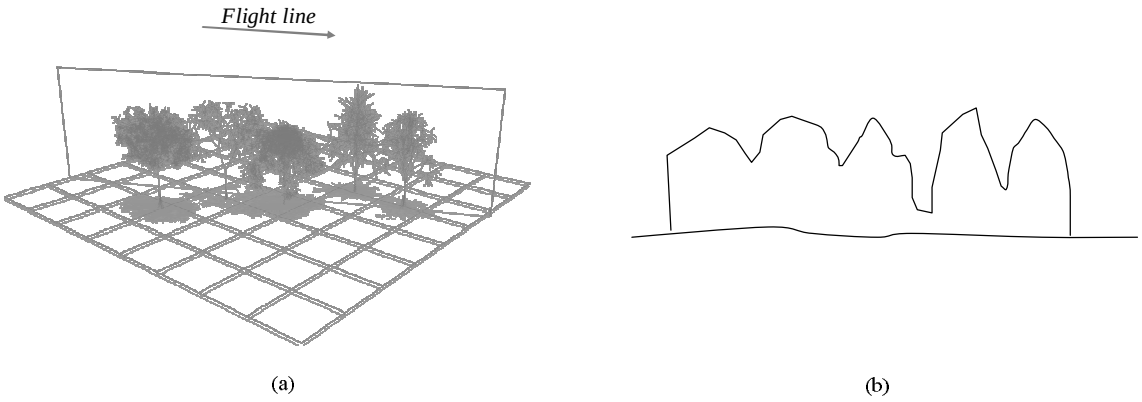


Figure 3.1. An example of a forest canopy profile which might be obtained via a laser profiling instrument.

To infer information on the area-based variables, LIS is needed to analyze profiling laser measurements. However, before applying LIS, the profiling laser data should be first processed and converted, preferably by some automatic data processing algorithms, to information relevant to the forest stand characteristics of interest. These processing algorithms includes those for reconstructing ground topography, deriving canopy height profiles, and stratifying the canopy height profiles into segments with respect to the types of land cover., The stratification of profiling transects is central to this study for the purpose of estimating percentage cover of each land cover. To stratify a laser height profile, advanced signal processing techniques such as wavelet analysis are effective in designing automatic algorithms. No elaboration on these aspects is made in this paper, and details concerning the implementation of relevant algorithms may appear in subsequent publications.

Airborne profiling measurements have been used in conjunction with LIS sampling techniques to inventory forest resources of large forested areas. Nelson et al. (2003b) used over 1300 km of systematic profiling transect measurements acquired on 14 flight lines 4 km apart to conduct an LIS-based multiple resource inventory of Delaware. Their estimates of merchantable volume fell within 21% of US forest service estimates at the county level, and 1% statewide, and their estimates of total above-ground dry biomass estimates were within 22% of USFS estimates at the county level

and within 16% statewide. They also used profiling measurements to estimate the areal extent of impervious and open water area for three counties of Delaware, and found that results were comparable to estimates from other sources. Using the same dataset, Nelson et al. (2005) applied the LIS method to estimate the extent of Delmarva Squirrel Fox habitat. Although their results showed that airborne profiling laser measurements in conjunction with the LIS method offers an effective tool for regional forest inventory, the variances of the LIS estimates reported by them only are an approximation to the true variances. They also claimed that the variance estimator in their study gives an upward conservative estimation (Nelson et al., 2005).

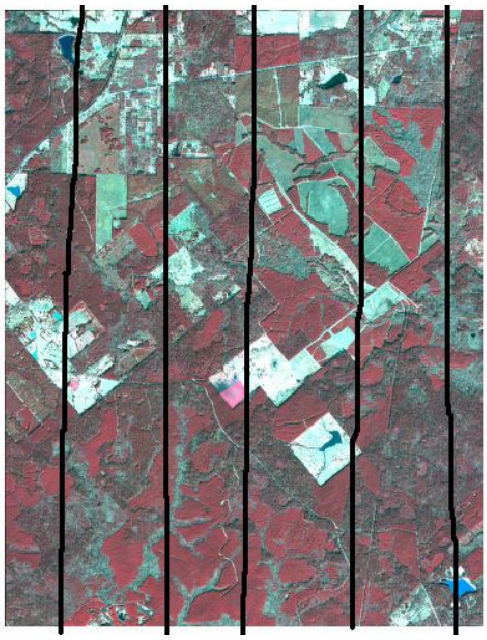


Figure 3.2. Parallel transects flown by a profiling LiDAR (profiler) over a study site: the dark lines are the flight lines of the laser profiler, the background is a false-color QuickBird image.

Because the variance or standard error of an estimate provides a quantitative measure of the uncertainty surrounding that estimate, reported means or totals without the associated estimates of variance are less valuable. Kaiser (1983) gave for his estimators a definition formula of the variance which includes two components, i.e., the variance terms and the covariance terms; and he stated that even in simple cases the

formula is mathematically intractable due to the unknown form of the covariance components. He also suggested obtaining a number (m) of independently and identically distributed estimates based on m independently chosen transects of certain length L, then pooling these m estimates by arithmetic averaging as a final estimate and taking the sample variance as the estimated variance. In terms of the intractability of the formula for covariance, DeVries (1986) had the same argument, and he offered a similar formula to evaluate the variance in the cases of transects of varying length, where the transect length is used as weighting in the calculation. Nelson et al. (2005) restated a successive difference estimator based on the study of Lindeberg (1926), and envisioned the use of a large Monte Carlo experiment to obtain a valid estimate of the estimator's variance for the systematic airborne LiDAR sampling. To effectively analyze and interpret the estimated forest attributes from profiling LiDAR measurements, we attempt to investigate the uncertainty of LIS estimates by quantifying its variance, which will serve as a crucial step in appropriately analyzing and interpreting profiling LiDAR measurements.

Specifically, this study will use both a computer-simulated artificial landscape and a real landscape classified from a Landsat ETM+ image in order to simulate transect observations in a systematic manner. Based on a series of simulated observations, four estimators for cover percentage of certain patch type are tested, and the corresponding variance estimators are also constructed and evaluated. To simplify the demonstration, we only considered binary landscapes: background versus patches of interest for the computer-generated landscape, and forest versus non-forest for the classified ETM+ landscape. For conveniences, we, throughout the paper, term 1-valued pixels as forestry and 0-valued ones as non-forestry in both the artificial and real landscapes. The results can be easily generalized to the multiple-type cases. Also, in this study, we formulate a theoretical framework for evaluating the variance estimators under certain assumptions.

3.3 Methods

3.3.1 Computer-simulated Landscape

A landscape simulation package SIMMAP, which implemented a modified random simulation method (Saura and Martínez-Millán, 2000), was utilized to generate a categorical landscape map with 2 patch types and an extent of 900 x 900 pixels (Figure 3.3). In the generation, the minimum patch size was set to 90 pixels. The forest cover percentage is the main landscape metric of inference interest. Because in practice, the shape of the study area, i.e., a county or state, could be arbitrary, a irregular-shaped mask was arbitrarily delineated on-screen in ENVI (ITT Visual Information Solutions, Inc.) and applied to the 900 x 900 image to clip out a region of interest, for the purpose of mimicking reality (Figure 3.3).

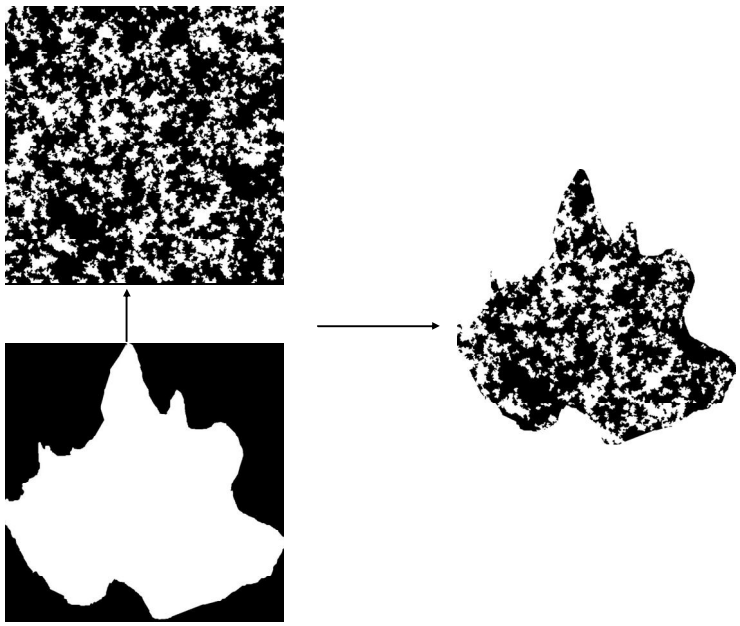


Figure 3.3. The computer-generated categorical map and a mask of the boundary of study area, and the resulting landscape after applying the mask to the original square region.

3.3.2 Classified Map from Remotely Sensed Data

A Landsat ETM+ multispectral image over Huntsville, East Texas, USA, was classified with the unsupervised ISO-DATA algorithm into seven classes. Then, the seven classes were grouped into two categories: forest and non-forest. Because the inference of interest is not the absolutely true value of the percentage covers for the Huntsville area, no accuracy assessment was performed. As with the artificial landscape, a mask was applied to the ETM+ landscape to mimic a natural shape of study area, and the resulting landscape is depicted in Figure 3.4.

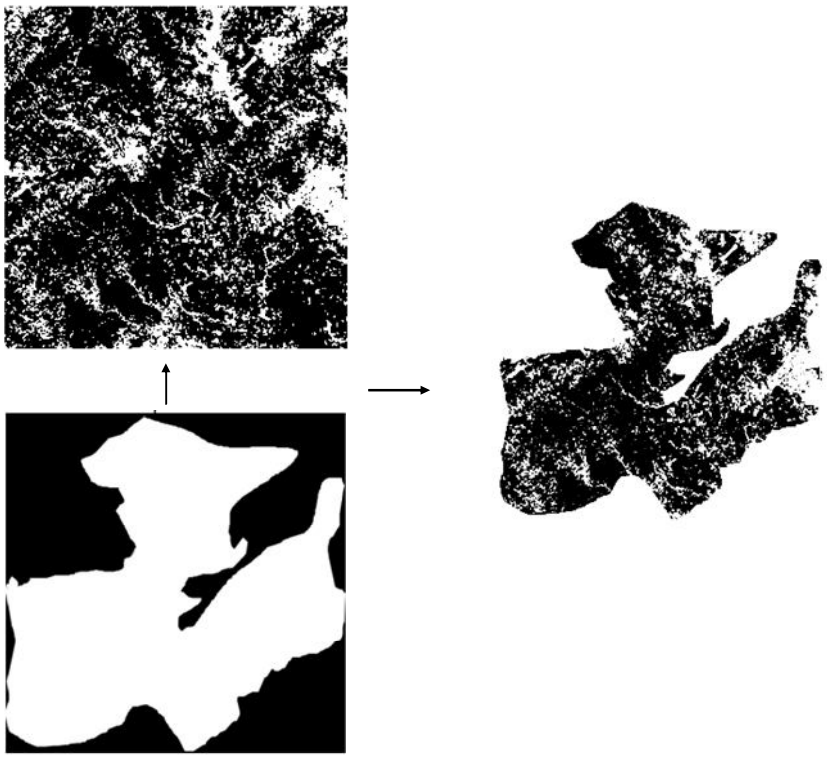


Figure 3.4. The classified Landsat ETM+ image with two classes (green for forest and black for non-forest), and a mask of the study area boundary, and the resulting landscape after applying the mask to the original square region.

3.3.3 Total Area of a Given Class

To simplify the illustration, it is assumed that the landscape contains only two categorical classes, i.e., non-forest (0) and forest (1). The interest is in estimating the total forested area. In practice, when a profiler system is flown, the flight direction is typically set and fixed; the sampling transects are parallel and spaced almost equally. A coordinate system was chose such that the *x* axis is perpendicular to the direction of sampling transects and the *y* axis is selected as the tangent to the left-most boundary (Figure 3.5). For a given transect at *x*, its length *l(x)* is equal to the distance between the upper and lower boundary points connecting the transect, excluding the portion that may lie outside the study area if any (i.e., the second transect from the left in Figure 3.5); and the proportion of the transact intercepted by forest patches is denoted by *r(x)*. Note that a transect may intercept many patches. Apparently, the total area of all forest patches *a* can be evaluated by,

$$a = \int_0^s r(x)l(x)dx \qquad (3.1)$$

where *s* represents the span of study area in the *x* direction.

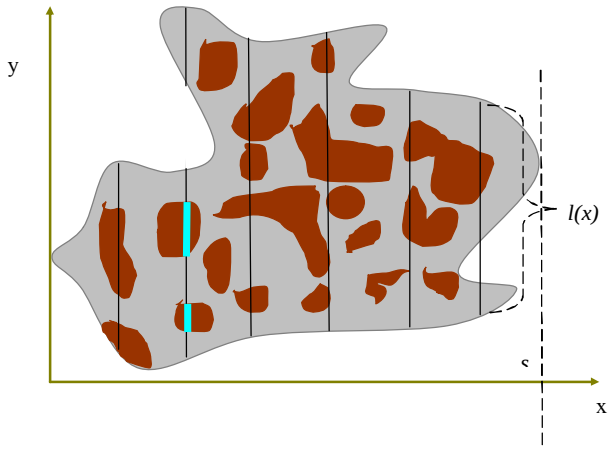


Figure 3.5. A hypothetical region of study where the black solid lines are transects for LIS and will be used to infer information on the colored patches, i.e., the cover percentage. The transect at x touches the upper and lower boundary points of the region, and has a length of l(x); the transect intercepts with patches, and denote the proportion intercepted to its total length is r(x).

To clarify notations, lowercase letters will denote variables or functions that are deterministic or fixed, while the uppercase is used to denote a random variable or function.

3.3.4 LIS Estimators for an Unknown Yet Fixed Landscape

The geographical extent of a study often is known by referring to some archived data source, e.g., a GIS layer. In addition, suppose that the orientation of all sampling transects are in the same direction and already determined as is the case for a profiling LiDAR survey. Hence, *l(x)* is known and can be calculated. The proportion of transect intercepted by the patches of a certain type (e.g., forest), $r(x)$, is fixed at any given *x*. Over the study area, $r(x)$ remains unknown except along transects where observations have been made. If the placement of transects is uniformly-random over the span [0, s] in the *x* direction and the property of $r(x)$ is unknown, the following estimator $\hat{a}_1$ can be effective:

$$a = s\int_0^s r(x)l(x)\frac{1}{s}dx \;\Rightarrow\; \hat{a}_1 = s\frac{\sum_{i=1}^{n} r(x_i)l(x_i)}{n} \qquad (3.2)$$

where *n* is the number of transects observed, and the variance of $\hat{a}_1$ is given by,

$$Var[\hat{a}_1] = \frac{s^2}{n}Var_{\frac{1}{s}}[r(X)l(X)] = \frac{1}{n}\left[s\int r^2(x)l^2(x)dx - \left(\int r(x)l(x)dx\right)^2 \right] \qquad (3.3)$$

where $Var_{1/s}[r(X)l(X)]$ means the variance of $r(X)l(X)$ when assuming *X* is uniformly distributed, i.e., $X \sim 1/s$. Since $Var[\hat{a}_1]$ requires evaluating $r(x)$ over [0, s], this cannot be used directly because $r(x)$ are not available for all *x*. Alternatively, the sample values of $r(x)$ at the observed transect, i.e., $r(x_i), i = 1...n$, should be substituted to calculate the sample variance of $r(X)l(X)$ as the estimated variance, that is,

$$\hat{Var}[\hat{a}_1] \approx \frac{s^2 \sum_{i=1}^{n} [r(x_i)l(x_i) - \overline{r(x)l(x)}]^2}{n(n-1)} \tag{3.4}$$

where $\overline{r(x)l(x)} = \sum [r(x_i)l(x_i)]/n$ is simply the sample mean.

Intuitively, the longer *l(x)*, the more contribution of it to the total area. Rather than being randomly placed, transects can be located by the probability of *l(x)/m*, where *m* is the total area of the whole study region. This means that there are more transects placed over *x*'s with larger *l(x)*. According to such a design, a natural estimator should be,

$$a = m\int_0^s r(x)\frac{l(x)}{m}dx \;\Rightarrow\; \hat{a}_2 = m\frac{\sum_{i=1}^{n} r(x_i)}{n} \tag{3.5}$$

where again, *m* is the total area of the whole landscape such that $m = \int_0^s l(x)dx$ and the corresponding variance will be,

$$Var[\hat{a}_2] = \frac{m^2}{n} Var_{\frac{l(x)}{m}}[r(X)] = \frac{1}{n}\left[m\int r^2(x)l(x)dx - \left(\int r(x)l(x)dx\right)^2 \right] \tag{3.6}$$

where $Var_{l(x)/m}[r(X)]$ is the variance of *r(X)* by assuming $X \sim l(x)/m$. Similarly, the observed samples $r(x_i)$ can be used to obtain an approximation to the true variance, and it follows that,

$$\hat{Var}[\hat{a}_2] \approx \frac{m^2 \sum_{i=1}^{n} [r(x_i) - \overline{r(x)}]^2}{n(n-1)} \tag{3.7}$$

3.3.5 LIS Estimators for a Random Landscape

In contrast to the aforementioned setting, the configuration of a landscape, i.e., *r(x)*, is supposed to be an unknown and random function of *x* and was denotes as *R(x)*. Consequently, the total area of forest cover (patch type of interest) is a random variable instead of an unknown constant. In reference to Eq. 3.1, it can be computed by,

$$A = \int_0^2 R(x)l(x) \;\Rightarrow\; \mu_A = E[A] = \int_0^s \mu_R(x)l(x)dx \tag{3.8}$$

where $\mu_R(x)$ is the mean function of $R(x)$, and in most cases, it may be not a constant function. Of particular note is that mathematically $R(x)$ cannot be a Gaussian random function because for any x, $R(x)$ is bounded over [0, 1]; however, for practical convenience, $R(x)$ may be approximately gaussian.

In this random setting of landscape configuration, the corresponding parameter of interest is $E[A]$, i.e., the mean of A. To simplify, assuming that $R(x)$ has a constant mean μ_R independent of x, then, $E[A]$ and $Var[A]$ then can be obtained respectively by:

$$\mu_A = \mu_R \int_0^s l(x)dx = m\mu_R \tag{3.9}$$

$$Var(\mu_A) = \int_0^s \int_{-x_1}^{s-x_1} C_R(x) l(x+x_1) l(x_1) dx dx_1 \tag{3.10}$$

where $C_R(x)$ is the covariance function of $R(x)$, and Eq. 3.10 holds if and only if $R(x)$ is widely-sense stationary, which should not be too rigid a requirement in most cases. By further assuming the mean-ergodicity of $R(x)$, a reasonable estimator of μ_A can be expressed as,

$$\hat{\mu}_{A,1} = m\hat{\mu}_R = \frac{m\sum_{i=1}^{n} R(x_i)}{n} \tag{3.11}$$

and its variance is,

$$Var[\hat{\mu}_{A,1}] = \frac{m^2}{n^2} 1' \cdot \Sigma_R \cdot 1 \tag{3.12}$$

where again, m is the area of whole landscape; $R(x_i)$'s are the sampled value at n transects that can be either systematically or randomly placed; Σ_R is the variance-covariance matrix of $R(x_i)$ and it can be constructed from $C_R(x)$ which, for the (i, j) entry of Σ_R, gives $C_R(|x_i - x_j|)$. Unfortunately, the structure of $C_R(x)$ is usually unknown; therefore, in practice prior knowledge about $C_R(x)$ may be entailed.

Another estimator of μ_A is given in Eq. 3.13 with Eq. 3.14 being the associated variance:

$$\hat{\mu}_{A,2} = s\frac{\sum_{i=1}^{n} R(x_i)l(x_i)}{n} \tag{3.13}$$

$$Var[\hat{\mu}_{A,2}] = \frac{s^2}{n^2}[l(x_1), l(x_2), ..., l(x_n)]\Sigma_R[l(x_1), l(x_2), ..., l(x_n)]' \tag{3.14}$$

The unbiasedness of $\hat{\mu}_{A,2}$ depends on both the properties of $R(x)$ and the sampling schemes of x_i. For instance, if $R(x)$ assumes a constant mean, $\hat{\mu}_{A,2}$ is unbiased when x_i's are randomly sampled. Conversely, suppose that $\mu_R(x)$ is dependent on *x*, $\hat{\mu}_{A,2}$ is biased when transects are observed at fixed locations x_i's.

3.3.6 Simulation Schemes

Systematic samples of transects are drawn randomly in a manner similar to that a profiling LiDAR observes. The spacing between adjacent transects are purposely designed to be randomly distributed around the average spacing since the flight lines of profiling LiDAR are not exactly equally spaced. Specifically, for both the computer-generated landscape and classified categorical map, a range of transect numbers are used; and at a given number of transects, a total of 30 realizations are simulated. The sample variance calculated from each group of 30 realizations is used to be compared with those obtained by the proposed variance formula.

3.4 Results

The proportion of transects intercepted by forest patches, i.e., $r(x)$, is plotted against *x* in Figure 3.6a and 3.6b for the artificial landscape generated by computer and the real landscape created from ETM+ imagery, respectively. The curves suggested that in both cases, the $r(x)$ appears to be randomly fluctuating and there exists some spatial correlation to a certain degree as also shown in Figure 3.7a and 3.7b which show the sample covariance function calculated from the corresponding $r(x)$'s. In both cases, the correlation monotonically decreases as the lag increases. These two sample covariance

functions as $C_R(x)$ were plugged into the formula in Eqs. 3.10, 3.12, and 3.14 developed above to compute the sample variance of the estimators.

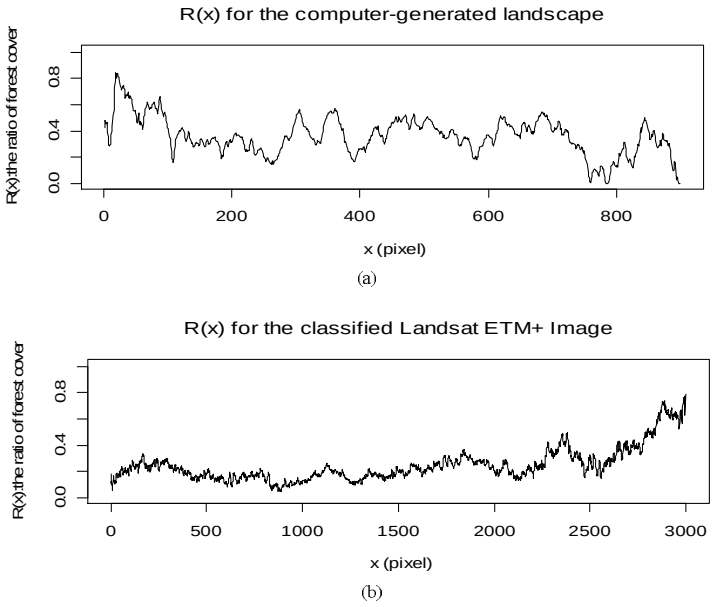


Figure 3.6. The proportion of transects intercepted by forest patches , r(x), is plotted against x (pixels) for the artificial landscape generated by computer (Fig. 3.5) and the real landscape created from ETM+ imagery (Fig. 3.4).

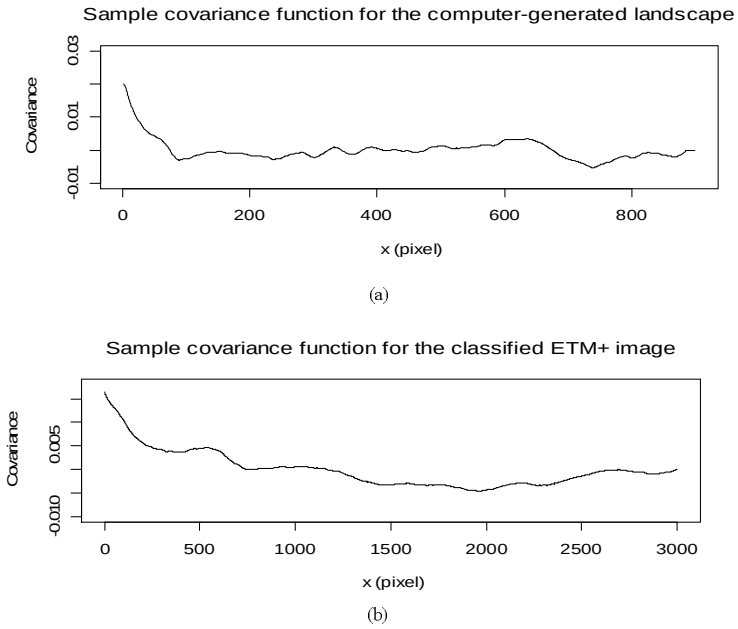


Figure 3.7. The sample covariance function calculated from the r(x)'s of Figure 3.5, respectively for the artificial landscape (a), and the classified land cover map (b).

Figure 3.8 shows the forest cover percentage estimated by four different estimators, at different choices of transect number. The forest cover percentage is simply the ratio of estimated forest area to the total area. The four estimators are respectively the one used by Nelson et al. (2005), and three of the four discussed above except the second one. They are denoted by "Nelson" or "N", "a1", "A1" and "A2" as labeled in the legends of Figures 3.8-3.11, respectively. The plotted values of estimates in Figures 3.8 and 3.10 only represent a random realization out of 30 ones. The true reference value of forest percentages are 0.368 and 0.228 for the simulated artificial and classified real landscape maps, respectively. In both cases, the estimated values tend to come closer to the true value as more transects are observed. Moreover, an overall trend of overestimation is observed for all the four estimators (Figure 3.8a) in the case of simulated landscape, although the estimates are occasionally lower than the true value, and the estimator of "N" and "A1" seem to consistently give the estimation closest to the true value. On the other hand, an opposite trend was noticed for the classified ETM+

image where overall, all four estimators were apt to underestimate the percentage cover.

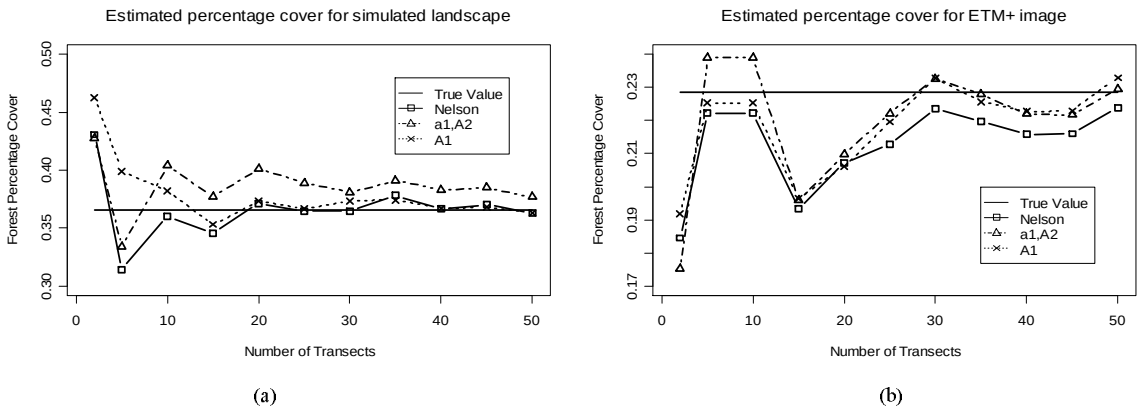


Figure 3.8. The forest cover percentage estimated by four different estimators as a function of the number of transects used. The four estimators are the ones used by Nelson et al. (2005), the first estimator a1 discussed in "Estimators when the configuration of landscape is fixed yet unknown", and the two estimators A1 and A2 discussed in "Estimators when the configuration of landscape is supposed to be random". (a) is the case for artificial landscape, and (b) for the classified remote sensing imagery.

One way of calculating estimator variance is to take the sample variance of 30 realizations, i.e., standard practice that is quite often employed in Monte Carlo simulation. The sample variances obtained so are supposed to offer a quite reasonable approximation to the true values inherent in the sampling design. Figure 3.9 suggests that for both landscapes the variances of four estimators were similar. To be more precise, for the simulated landscape, the estimated standard errors (SE) from the simulation with "Nelson" estimator is the lowest with "a1" and "A2" providing the highest estimates, while for the classified image, the estimated SE of "Nelson" and those of "a1" and "A2" are very close especially at a large number of transects. In addition, another way to calculate the estimator variance is to directly use the aforementioned formula which is derived under certain assumptions, with results given in Figure 3.10. As with the Monte Carlo simulated estimates, the SEs calculated with variance formula appear to follow a similar trend except that "a1" and "A2" have different estimates due

to the difference in assumptions. Moreover, in contrast to the decreasing variance with more transects for Nelson's estimator and a1, A1 and A2 exhibit an almost invariant pattern.

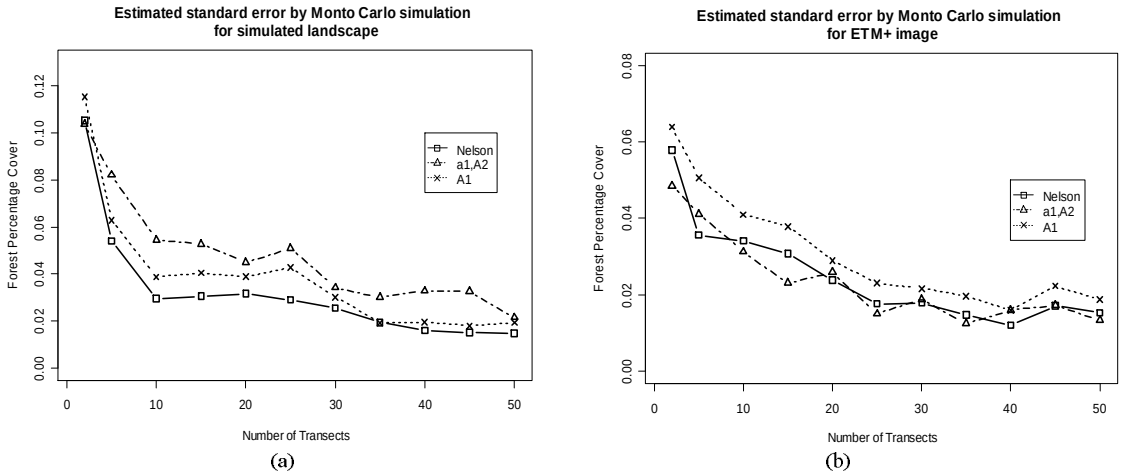


Figure 3.9. The estimated variance by sample variance of the simulation: (a) is the case for artificial landscape, and (b) for the classified remote sensing imagery

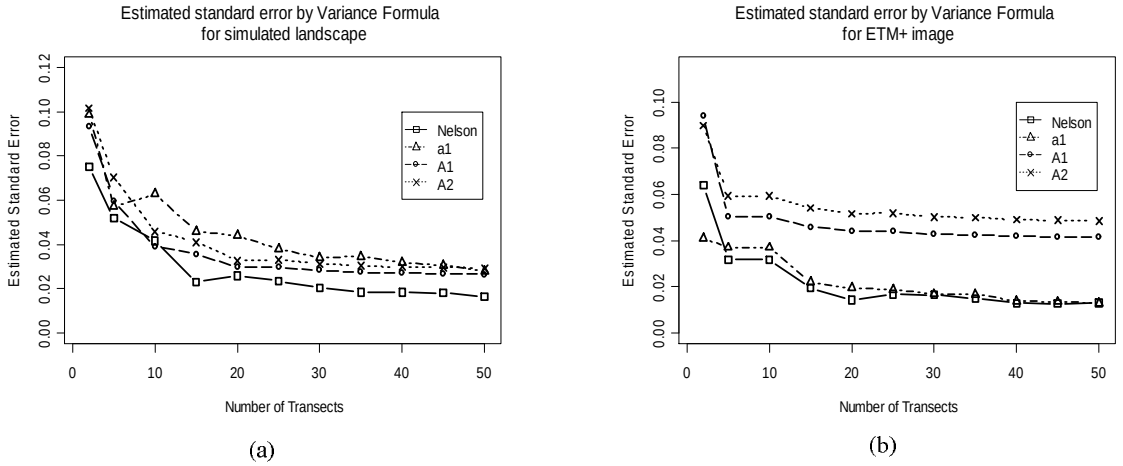


Figure 3.10. The estimated variance calculated by the variance formula: (a) is the case for artificial landscape, and (b) for the classified remote sensing imagery

To further compare the two methods of computing variance, the ratios of SEs derived from the formula to that from the Monte Carlo simulation are presented in Figure 3.11. Assuming that the simulation-derived sample variance is a more reasonable estimate, the closer to 1 the ratio is, the more valid the formula for variances is, or a more reasonable assumption is made as to $r(x)$. As shown in Figure 3.11, the resulting plots reveal that in the case of simulated landscape, the ratios fluctuated around 1.0 for all four variance estimators, e.g., with a range from 0.72 to 1.41 for the Nelson's estimator; However, for the real landscape of classified image, "A1" and "A2" provided inflated estimates compared to the reference values from the Monte Carlo simulation, with "A2" producing the most upward estimates and "A1" following next. In contrast, Nelson's estimator and "a1" have fluctuating ratios that are more closer to 1, which suggests that in this case these two estimators offer more realistic estimates of variance.

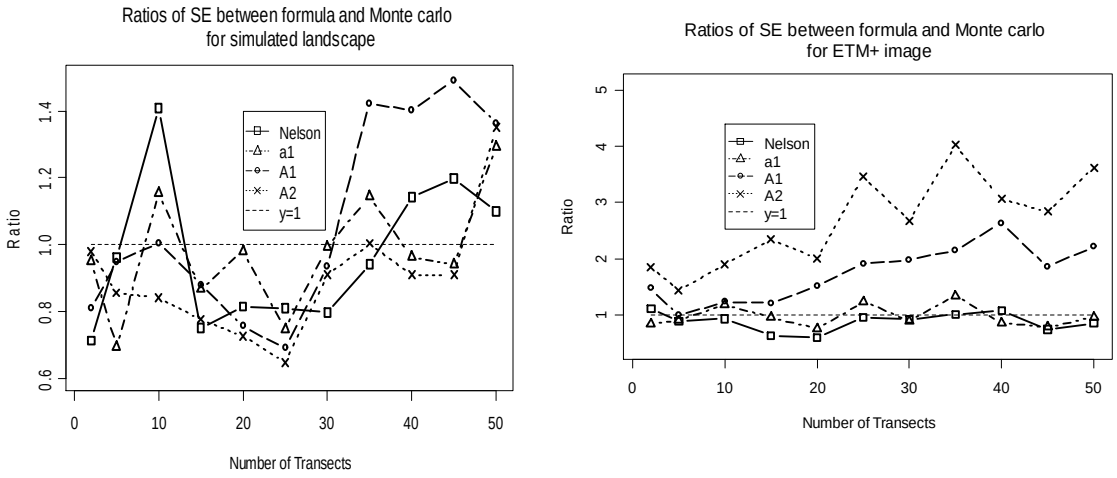


Figure 3.11. The ratio of the formula-derived variance (shown in Figure 3.9) to the sample variance of simulation (shown in Figure 3.8) as a function of number of transects used. The closer to 1 the ratio is, the more valid the variance formula is. (a) is the case for artificial landscape, and (b) for the classified remote sensing imagery.

3.5 Discussion and Conclusions

By mimicking the way a profiling LiDAR collects linear transect observations, the simulation of random realizations makes it possible to examine the variability of simulated samples as well as quantify the variance of a given estimator. In this study, estimators for the percentage cover of a given class is tested; more importantly, the variance of these estimates are obtained by either taking sampling variance of the simulation or using the proposed formula as expressed in Eqs. 3.4, 3.7, 3.12 and 3.14.

Two settings of landscapes, whether deterministic or random, are assumed for deriving both the estimator and variance formula. Instead of referring to geometrical probability as used in Kaiser (1983), Monte Carlo integration approach is employed for developing estimators of percentage cover in this study; and the basic ideas are illustrated by Eqs. 3.1, 3.2, 3.5 and 3.8. The estimator ($\hat{a}_1$ or a1) of the first setting appear to be quite similar to the one used by Nelson et al. (2005), but differences between them exist that $\hat{a}_1$ is an estimator for the area of patches of interest that needs to be divided by the already known value of total area of the study area to obtain percentage cover, while the estimator in Nelson et al. (2005) provides a direct estimate of percentage cover. The estimators for the random setting, however, tend to have a much more inflated variance that seems much larger than those obtained by the Monte Carlo simulation. This inflation may be caused by treating the systematical sampling as random when evaluating the variance, and it may also be an artifact of introducing randomness to depict the fixed but unknown landscape, i.e., *r(x)*. Moreover, the inference in the second setting of random lanscape can be only made as to the mean value since *R(x)* is random by itself. As such, the simulation of drawing transects from a fixed landscape is not too methodologically rigorous; but this appeared not to be a problem as demonstrated by the results for the simulated landscape.

In fact, the apparent disparity between the formula-based and simulation-based variance estimates for “A1” and “A2”, as indicated in Figure 3.11. b for the classified image, can be most likely attributed to the departure of *R(x)* from the underlying

assumptions made for deriving the variance estimators. It becomes clear that a drift is present in *r(x)* as depicted in Figure 3.6b, which shows a high tail at large *x*. As a result, the incorporation of such an overall trend into the calculation may mislead the estimation of covariance function as shown in Figure 3.7b, since covariance function is supposed to capture only the random components of underlying stochastic process. Consequently, the inappropriate knowledge about covariance function will result in unreasonable variance estimates, as also noted for the variance estimator in Kaiser (1983). In this study, estimates of covariance function were computed from $r(x)$. However, in practice, $r(x)$ remains unknown and is of inference interest based on the available samples observed at transects. Because of the limited number of transects, it seems impractical to estimate the covariance function, thus preventing the effective use of variance estimators formulated for the random setting. A possible remedy is to fit a parametric covariance model of known form to the available observation; specifically, the parameter relevant to the magnitude is simply estimated by the sample variance of $r(x_i)$'s, and parameters relevant to effective correlation distance may be inferred by analyzing the observations along the transect direction (i.e., the y axis in Figure 3.5), due to continuous measurements available in this direction. However, the development of an efficient and reliable fitting procedure needs further investigation in future studies.

In our derivation, it is assumed that the geometry of study areas is known, which means that given a direction, $l(x)$ can be calculated. This is typically the case given that the study area is often digitally characterized in a GIS. By changing the observation direction, not only $l(x)$ but also $r(x)$ changes; therefore, the orientation of transects has great influence on the properties of relevant estimation. There also may exist a favorable direction for high-precision estimates; for example, an extreme scenario is hypothesized in Figure 3.12 where the dotted half of the square region represents forested area. If a single vertical transect is randomly observed (the left of Figure 3.12), the sampled percentage can take either 1 or 0 with equal probabilities; however, when observed horizontally, the true value 1/2 is always observed. To this end, prior knowledge, if available, can be used to guide the selection of appropriate direction; otherwise, the

direction that could minimize the variability of $l(x)$ may be an option. In fact, when $l(x)$ is constant, all the four estimators as well as the one used in Nelson et al. (2005) essentially produce the same estimates, but they do have different variance estimates due to different assumptions. In deriving the LIS estimates or the corresponding variance estimator such as $\hat{a}_1$ and $\hat{\mu}_2$, only the transect length at observed locations is used. To compensate this, the primary author suspects that incorporating all information on *l(x)* should increase the precision of estimates.

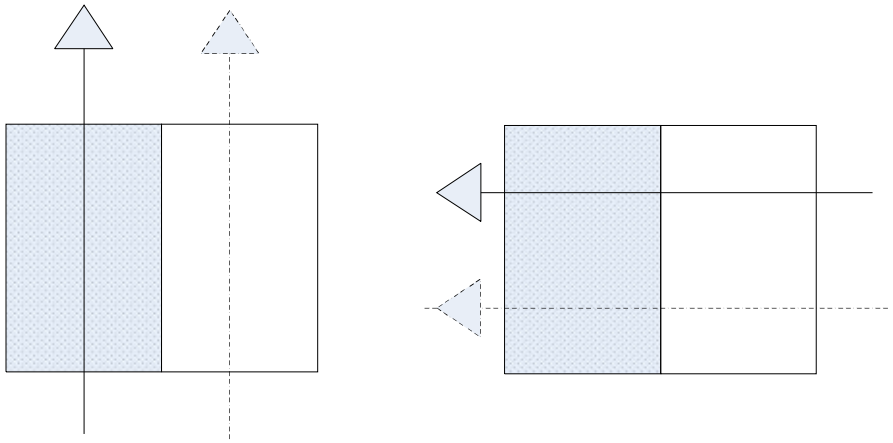


Figure 3.12. An extreme hypothetical landscape of square shape where the dotted half represents forested area, for the purpose of demonstrating the effects of transect orientation : vertically-placed transects (the left) and horizontally-placed transects (the right)

As a final note, although the aforementioned derivation is based on binary categorical landscapes, the proposed estimators can be equally applied to estimate percentage cover for landscapes with multiple categories without any modification. Moreover, they can also be extended for estimating the mean value of a variable that varies continuously across the study area, e.g., regional biomass. To illustrate this point, suppose that the variable is a spatially-explicit function denoted by $f(x, y)$; then, by assuming the same geometry as in Figure 3.1, the mean of $f(x, y)$ averaged over the

study area is simply obtained by an integration as follows,

$$\bar{f} = \frac{1}{m}\int_0^s \int_{y_l(x)}^{y_u(x)} f(x,y)dydx = \frac{1}{m}\int_0^s f_y(x)l(x)dx \qquad (3.15)$$

where $f_y(x) = \int_{y_l(x)}^{y_u(x)} f(x,y)dy$; $y_l(x)$ and $y_u(x)$ are the y coordinates respectively for the lower and upper boundary points at *x*, and m is the size of study area. In Eq. 3.15, $f_y(x)$ plays the same role as the $r(x)$ in Eq. 3.1. Hence, after the samples of $f_y(x)$ are observed at a number of transects that are parallel to *y*, $\bar{f}$ can be estimated by using the same estimators as previously discussed.

The logic to formulate the estimators for LIS is a little different from the mostly used geometrical probability approach by Kaiser (1983). Hence, it still remains undetermined whether a logically sound reasoning line is followed to deal with the LIS problems in this study, or not. Nevertheless, the attempt to rely on Monte Carlo integration random simulation for evaluating LIS estimators points to a bright direction towards the ease of practical implementation of LIS schemes rather than the mathematical intractability.

CHAPTER IV
SCALE-INVARIANT PREDICTION OF FOREST BIOMASS USING AIRBORNE LASERS

4.1 Overview

Researchers in LiDAR (Light Detection And Ranging) have been striving to search for the most appropriate LiDAR-based predictors in regression models for estimating forest structural variables. Most of previously developed models are scale-dependent, and need to be fitted and then applied both at the same scale or cell size. The objective of this paper is to develop methods for scale-invariant prediction of forest biomass using LiDAR data. We proposed two scale-invariant biomass models: a linear model that uses LiDAR-based canopy height distributions (CHD) as predictors, and an equivalent nonlinear model that uses the whole canopy height quantile function (CHQ) as predictors. The two models were justified by a mathematical framework with several moderate assumptions. The validation of models was based on a series of realistic plot-level data that have been synthesized from the LiDAR Canopy Height Model (for predictors) and LiDAR-derived detailed biomass map over a forested region in eastern Texas. The models were also applied to six transects of profiling laser data for estimating regional biomass. The results from intensive evaluations show that the proposed models can accurately predict biomass and have consistent predictive performances across a variety of scales with an R^2 ranging from 0.80 to 0.95 among all the fitted models. The results also show that in our experiments, a training sample size of around 50 plots or less is enough to guarantee a good fitting of the linear model. The findings of this study suggest that the proposed models are effective for forest inventory tasks where the analysis units vary in size and shape, and they also hold promises for predicting other forest structural variables such as basal area, timber volume, crown fuel weight, and Leaf Area Index.

4.2 Introduction

Accurate estimation of biomass is essential for a better understanding of carbon cycles over terrestrial ecosystems where forests serve as a primary reservoir of terrestrial carbon (Houghton, 2005). Monitoring changes in biomass over time has both environmental and economical significance as typified by the establishment of the Kyoto Protocol. In the context of global change and warming, quantification of carbon sequestration in forested areas provides insights into relevant biogeochemical processes, e.g., vegetation responses to elevated CO_2 levels, environmental impacts of anthropogenic activities, and characterization of carbon sources and sinks at regional or global scales (Houghton et al., 2001). A growing need for spatially-explicit mapping of forest biomass has been partially compensated by recent advances in remote sensing technologies. In particular, Light Detection And Ranging (LiADR) recently emerged as a breakthrough technology in remotely inventorying forest resources. It features direct characterization of canopy structures, which makes it a superior choice for accurate estimation of above-ground biomass (AGBM) as compared to optical remote sensing that suffers from saturation of spectral response to dense canopies with high biomass (Lefsky et al., 2002).

Small-footprint airborne laser scanners (ALS) with discrete-returns continuously transmit laser pluses to pinpoint the intercepted targets by accurately timing the roundtrip of a pulse between the sensor and a target. LiDAR data from ALS, thus, consist of a collection of spatially-distributed points each represented by a coordinate triple of (x, y, z) and often tagged with such auxiliary information as reflected intensity, scanning-angle, and acquisition time. The mainstream of LiDAR forestry applications thus far more appreciates the use of coordinate information of triples than LiDAR auxiliary variables, with only few studies partly investigating the applicability of the auxiliary information. (Brandtberg et al., 2003; Lim et al., 2003; Holmgren, 2004; Popescu and Zhao, 2008). Previous efforts demonstrate the potential of LiDAR for extraction of forest structural attributes at different levels (Næsset, 1997; Omasa et al.,

2003; Nelson et al., 2003a; Popescu, 2007). A variety of algorithms were proposed to delineate single trees from LiDAR data and estimate tree height and/or crown dimensions at individual tree levels; most results suggest LiDARs tend to underestimate tree height due to a large probability of missing treetops even with a high sampling density (Popescu et al., 2002; Yu et al., 2004; Chen et al., 2006). Plot- or stand-level estimates of forest characteristics from LiDAR, e.g., mean tree height, basal area, AGBM, and timber volume, have also been reported for various forest types and conditions (Riaño et al., 2004; Næsset, 2004; Popescu et al., 2004; Andersen et al., 2005; Maltamo et al., 2006; Mutlu et al., 2008).Standard practice in establishing LiDAR prediction models for plot-level forest attributes involves relating the spatially coincident in-situ measurements, which are usually prorated if not temporally concomitant with laser data, to some carefully-selected LiDAR metrics by regression analysis. Upon validation, these regressed models will be applied on the rest of LiDAR data for prediction purposes (Nelson et al., 1988; Næsset and Bjerknes, 2001; Næsset, 2002).

The use of LiDAR for AGBM generally follows the same aforementioned two-stage procedure where ground reference biomass is obtained with destructive sampling, or more often with the recourse to allometric equations. Previous studies showed promising results in estimating biomass with LIDARs; however, the selected LiDAR metrics and regression models often lack commonalities as also noted by Lim and Treitz (2004). The reported LiDAR metrics mainly include mean, maximum and median canopy height, quadratic mean canopy height, quantile heights and etc.; these metrics were used alone or combined in linear models or nonlinear models. (Lefsky et al., 1999; Means et al.,1999; Nelson et al., 2004). Nelson et al. (1988) found, for example, that the logarithmic model with mean height of all laser returns as predictor accounts for the most variation in ground-measured AGBM using first-return laser profiler data. Popescu et al. (2003) reported two plot-level linear biomass models for pine and deciduous plots respectively; the significant predictors of the pine model include the LiDAR-derived mean and maximum crown diameter, and those of the deciduous model include the LiDAR-derived maximum tree height and mean crown diameter. Also, with laser

scanner data, Lim et al. (2003) exploited three linearly-transformed single-predictor models that use the mode of laser returns, the mean height of all laser returns, and the mean height of only laser returns with intensity values above a preset threshold, respectively. In another study, Lim and Treitz (2004) explored laser canopy-based quantiles for biomass by assuming that the laser canopy height distribution (CHD) is a function of the vertical distribution of leaf areas, and they concluded that all quantiles have similar prediction capabilities if the tree allometry in forests remains the same.

The importance of scale in remotely sensing surface biophysical parameters has long been recognized, partly due to the incapability of sensors to directly measure the heterogeneity at sub-pixel scales (Strahler et al., 1986; Woodcock and Strahler, 1987; Marceau, 1999). From either an applied or philosophical perspective, scale issues arise virtually from all disciplines of the earth sciences (Oreskes et al, 1994). As far as remote sensing theories and applications are concerned, scale problems often reveal themselves by a range of incompatibilities among relevant scales such as those at which sensors take measurements, those at which physical variables of interest are defined in theory and observed *in-situ*, those at which remote sensing application models are formulated and applied, and those at which retrieval algorithms together with the involved remote sensing forward and inversion models are formulated and implemented; moreover, ad-hoc image processing procedures such as resampling and filtering often introduce compounding scale effects (Turner et al, 1989; Marceau et al., 1994; Marceau, 1999). To compensate those scale discrepancies, practical applications often require devising some viable scaling-up or scaling-down schemes. In addition, for a specific application, e.g., tree species mapping (Marceau et al., 1994), there often exists an optimal spatial resolution that may not be valid to other applications. In the context of LiDAR remote sensing of forest structures, most previous studies built prediction models that are likely to be not only study- or species-specific but also scale-dependent (Lim and Treitz, 2004; Patenaude et al., 2004).These models are typically trained using data collected at a given plot size, e.g., the field-measured biomass versus the LiDAR quadratic mean height, and for prediction purpose, they need to be applied at a scale or pixel size commensurate to

the plot size used in model training (Næsset, 2002). For example, when applying the regression equations trained at a plot size of 200 m^2 to predict stand-level forest characteristics, Næsset (2002) divided each stand into cells of 200 m^2 and averaged the predicted values of individual cells within a stand for obtaining the overall prediction of that stand. Noteworthily, a possible concern arises that some model variables involved, either predictors (LiDAR metrics) or dependant variables (forest attributes), are non-additive, by which we mean that the average of individual value of a variable at many cells is not equal to its value over the area aggregated from these cells with regards to the definition of that variable. In fact, the scale-dependence of LiDAR approaches was explicitly reported in a ground carbon study where the authors found that using the same exponential model with the 80^{th} LiDAR height percentile as the predictor, the total AGBM calculated by applying the model to each cell is not the same as the one obtained by using the single 80^{th} percentile of the whole area (Patenaude et al., 2004). They attributed the discrepancy to the nonlinearity of the exponential model, and suggested modeling height distributions across the stands, the woodland, or any area of interest as an alternative approach to the use of percentiles (Patenaude et al., 2004). To be more precise, even if the model is linear with a percentile height as the only predictor, this scaling effect still exists due to the nonadditivity of the percentile from two height distributions. In general, the scale-dependence of these models is revealed by a simple observation that the aggregation of prediction at individual cells is not the same as the single prediction based on the single predictor over the aggregated cells (Patenaude et al., 2004). This scaling effect may spring from one or more of the nonadditive properties of predictors, models, and responses when scale changes (Oreskes et al, 1994).

One way to quantify AGBM using LiDAR with minimal scaling effects is to inventory forests at individual tree levels and calculate biomass per tree (Popescu, 2007), and AGBM at scales above tree levels will be immediately available by integrating tree-level results up to the desired scale. Such a method involves the delineation of trees from LiDAR data. Computational demands by tree segmentation should be of little concern even for large areas considering the powerful capacities of modern computers. A

possible concern, however, is the incapability of current tree-segmentation algorithms to identify certain trees, e.g., understory and suppressed trees, as well as the inefficiency of the algorithms in delineating trees under certain canopy conditions such as dense forests and grouped trees (Maltamo et al., 2004). This limitation probably explains why little literature is found to estimate plot or stand-level AGBM through the use of tree-segmentation -and -integration approach, although there is literature reporting that dominant trees contribute significantly to the AGBM for certain forest conditions (Zavitkovski, 1976).

The main objective of this study is to explore the possibility of developing scale-invariant prediction models of ABGM from airborne laser scanner data. To this end, we proposed two scale-invariant models: One is a linear functional model that uses the whole curve of LiDAR-derived canopy height distribution (CHD) as the predictor, and another is a nonlinear model equivalent to the first one that instead uses canopy height quantile function (CHQ) as the predictor. We present a "naive" conceptual framework to provide theoretical justification for the development of our biomass models, and we also evaluate the models based on a series of synthesized training and testing data sets that were extracted through the combined uses of field measurements and LiDAR data over a forested area in eastern Texas, USA. Because the proposed models, once developed at a given scale, are supposed to make prediction for any forest inventory unit regardless of its size and shape, we applied the fitted model to six transects of profiler laser data collected over the same forested area by a Portable Airborne Laser System (PALS) both for evaluating the effectiveness of the biomass models and for assessing the accuracy of PALS estimate in reference to the scanner estimate.

4.3 Materials

4.3.1 Study Area

A 4800-ha forested region in eastern Texas, USA, was chosen for this study (Figure.4.1). The airborne laser coverage consists of pine plantations in various

developmental stages, old growth pine stands in the Sam Houston National Forest, many of which with a natural pine stand structure, and upland and bottomland hardwoods. Much of the southern U.S. is covered by forest types similar to the ones included in our study area, with similar forest types, productivity, and patterns of land use change. The study area is characterized by a gentle topography with a mean elevation of 85 m.

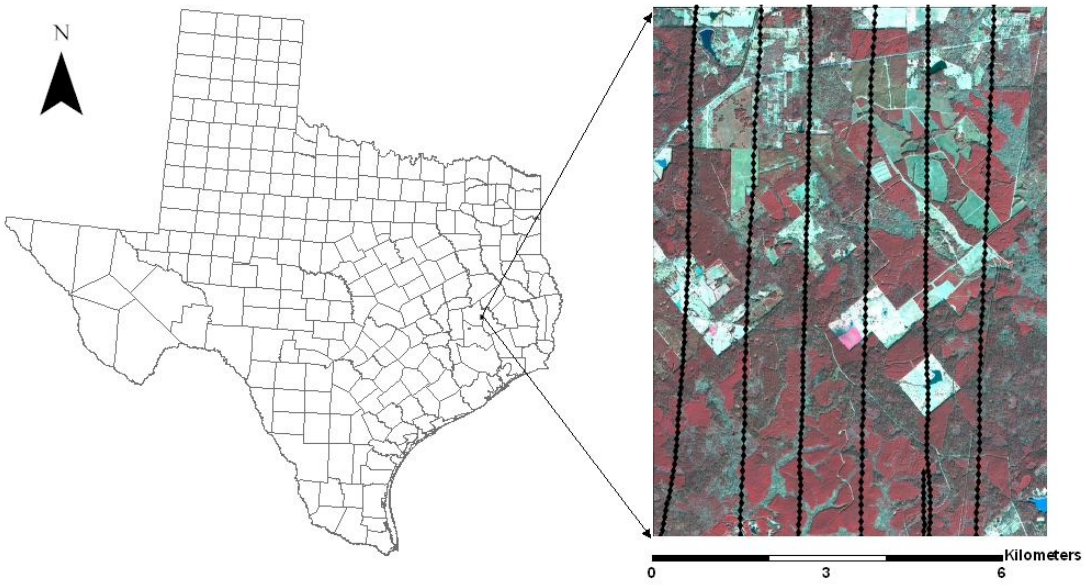


Figure 4.1. A map of Texas (left) with the study area (right) located in eastern Texas: The study area is a forested region depicted by the Quickbird Image, and the six vertical lines represent the profiling laser transects measured by the Portable Airborne Profiling Laser System (PALS).

4.3.2 Field Measurements

Fieldwork was undertaken during May-Jun. 2004 on 62 randomly selected circular plots, including 26 0.01-ha plots and 36 0.1-ha plots. The smaller plots are all within young unthinned pine plantations. A total of 1004 trees were tallied with respect to height, crown width, height to crown base, diameter at breast height (DBH), species, and crown class (Kraft). Tree height was measured using a Vertex Forester hypsometer;

DBH was measured using a diameter tape; crown width was calculated as the average of four perpendicular radii measured with a tape from the bole of a tree towards each cardinal direction, and crown class was determined as one of four categories, i.e., dominant, co-dominant, intermediate, and overtopped. (USDA Forest Service FIA National Core Field Guide, 2005, p. 78). The 261 overtopped trees are excluded in this study because they only intercepted a limited number of laser pulses and it is difficult to reconstruct their structures from LiDAR data. Of the 743 trees left, 504 are Loblolly pines (*Pinus taeda* L.) and 239 are deciduous trees, such as water oak (*Quercus nigra* L.), red oak (*Quercus falcata* Michx), sweetgum (*Liquidambar styraciflua* L.), and post oak (*Quercus stellata* Wangenh.).

Tree locations were indirectly calculated according to trigonometric relations using the geographical coordinates of plot centers and the azimuth and distance of tree boles relative to the plot center that have been measured by a Differential GPS, a Suunto compass (KB-14), and the Vertex hypsometer, respectively. Thus, the mapped tree locations refer to boles, not treetops that may deviate from tree boles. All of these factors caused difficulties in matching the field-measured trees with LiDAR-derived trees delineated by the tree segmentation algorithm. As a result, only 117 out of 743 measured trees were matched to LiDAR-derived trees including 94 pines and 23 deciduous trees. The matching was obtained with high confidence by on-screen examination on the overlaid layers of the field-measured trees, LiDAR-derived CHM, and LiDAR-derived trees (Popescu and Zhao, 2008). The descriptive statistics of the matched trees are summarized in Table 4.1.

Table 4.1. Descriptive Statistics of the field-measured trees that have been correctly matched with LiDAR-derived trees.

	DBH (cm)	Height (m)	Crown diameter (m)	Height to crown base (m)
Minimum	5.23	8.83	0.83	3.2
Maximum	78.49	37.49	13.27	25.9
Range	73.26	28.66	12.44	22.7
Standard deviation	17.21	6.91	2.68	4.45
Average	30.99	19.98	5.94	11.76

4.3.3 Airborne Laser Data and Multispectral Imagery

Airborne scanning LiDAR data were collected with a Leica-Geosystems ALS40 during the leaf-off season in March 2004 by M7 Visual Intelligence of Houston, Texas. The LiDAR system was operated to record two returns per pulse, i.e., first and last, with a reported accuracy of 20-30 cm and 15 cm for horizontal and vertical positioning, respectively, and was configured to scan +/- 10 degree from nadir, with an average swath of about 350 m wide on the ground. The dataset features a full coverage from either of two perpendicular directions, with 19 flight lines in the north-south direction and 28 in the east-west direction, resulting in an average of 2.6 laser hits per m^2. A Digital Elevation Model (DEM) derived from the LiDAR data using a proprietary package was also delivered by the data vendor.

The profiling LiDAR measurements were acquired in February 2004 using the PALS instrument developed at NASA (Nelson et al., 2003a). As a portable, light-weighted, and inexpensive system, PALS provides a feasible means for large-scale inventory forest resources, e.g., at statewide, national, and sub-continental scales. For this study, the PALS system, has a spacing of .25 m along the flight transect, flying at a altitude of 150-300 m above ground with a pulse rate of 2000 per sec. and a sample ratio of 10:1, and recording only first returns. The spatially coincident PALS transects within scanner-sampled area were spaced about 1 km apart and oriented approximately in the north-south direction. Since our proposed biomass model can be applied without any

modification to both scanner and profiler laser data as described later, a sub-objective of this study is to evaluate the accuracy of profiler biomass estimates, usually reported at a regional level, against the reference estimates of laser scanner data.

A QuickBird image over the study area was acquired in 2004. It was classified by the maximum likelihood method to differentiate between pines and deciduous trees whereas the other classes such as grounds, water bodies, and buildings, were grouped as a non-forest class. The classification yielded an overall accuracy of 89.7% that was obtained by examining 100 randomly generated pixels for each class.

4.4 Methodology

4.4.1 Overview of Methods

The primary purpose of this study is to formulate scale-invariant biomass models using airborne laser data. It involves selecting appropriate LiDAR metrics and model forms so that the models are developed at a given scale but can be applied to make prediction of biomass at the same or different scales with minimal scaling effects. The development and validation of such models require datasets of ground-reference biomass at continuous, or at least, multiple scales (or plot sizes). Such type of field data, however, is rarely available in most studies including this investigation. Therefore, we resorted to a LiDAR-based individual-tree segmentation approach to create a fine-resolution raster map of AGBM that serves as a reliable substitute to field data for synthesizing reference biomass values at a desired scale or plot size by properly aggregating the AGBM map to that scale. The synthesized plot-level AGBM data from the resulting map was used for training and testing two scale-invariant biomass models.

For clarification purposes, we state the major reasons why we prefer such an approach rather than directly using field measurements on ground plots:

(1) All the field plots are fixed in size and circular in shape, thus failing to provide ground-reference biomass at various scales or to study the effects of shapes of

analysis units, while the LiDAR-derived AGBM map allows for simulating the biomass observations at a variety of plot sizes or for different plot shapes.

(2) A considerable number of reference AGBM values can be generated at each scale from the LiDAR-derived biomass map, which, therefore, produces enough data points for model validation, as opposed to the limited number of field plots (i.e., 62 plots).

(3) LiDAR metrics that will be related to ABGM in the scale-invariant biomass model are extracted from the same raw LiDAR data source as the LiDAR-derived biomass map; this allows for seamless co-registration between independent and dependent variables, thereby, avoiding the mis-registration problem that usually is a concern when pairing field-geolocated plots with airborne-geolocated LiDAR CHM, particularly at individual tree levels.

(4) The data for training (i.e., calibration) and testing (i.e., validation) were generated in the same manner; this consistency makes model validation more viable because uncertainties caused by disparities of data sources are avoided.

Figure 4.2 outlines the key steps to create the reference AGBM map. These steps will be elucidated in the next sub-section. Following that, we develop a theoretical and methodological framework to build two equivalent scale-invariant biomass models. For the two models, we introduce a K-function as model parameters that need to be estimated from the synthesized training data. Predictors we chose for the two models are Canopy Height Distribution function (CHD) and Canopy Height Quantile function (CHQ), respectively. The models are considered as functional models because the predictors are themselves functions. Finally, we describe practical schemes for training models.

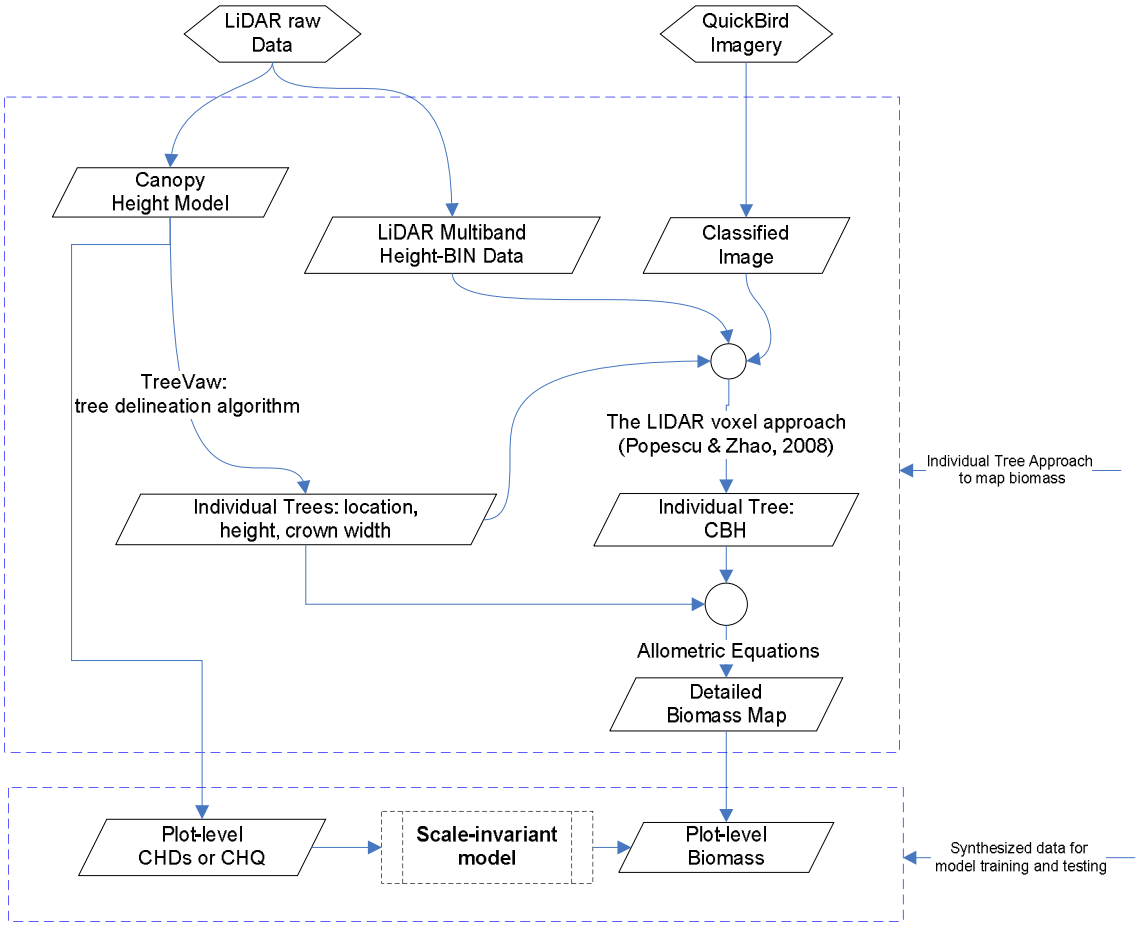


Figure 4.2. A flow chart summarizing the procedures employed to map individual tree biomass in order to simulate reference data for training and validating the proposed scale-invariant biomass models: In the synthesized data, the predictors are CHDs or CHQs at plots of given size, and the dependent variables are biomass at coincident plots.

4.4.2 Processing of PALS Profiling Laser Data

Each flight line of the PALS first-return data was viewed as a 1-D signal (Figure.4.3), and processed following the commonly reported procedures. First, ground hits were identified using the local minimum filtering, and then a spline was fitted to these selected ground hits to establish a ground curve. The conspicuous errors in the interpolated ground curve were eliminated by manually adding some potential ground

hits and deleting some mis-identified ones through on-screen visual examination. Afterwards, the canopy height profile was readily obtained by subtracting the ground curve from the top envelop of laser returns. Of particular note is that flight lines can potentially traverse across features such as roofs that are undesirable for forestry applications; therefore, segments falling on human-made building were isolated with the aid of the classified QuickBird imagery.

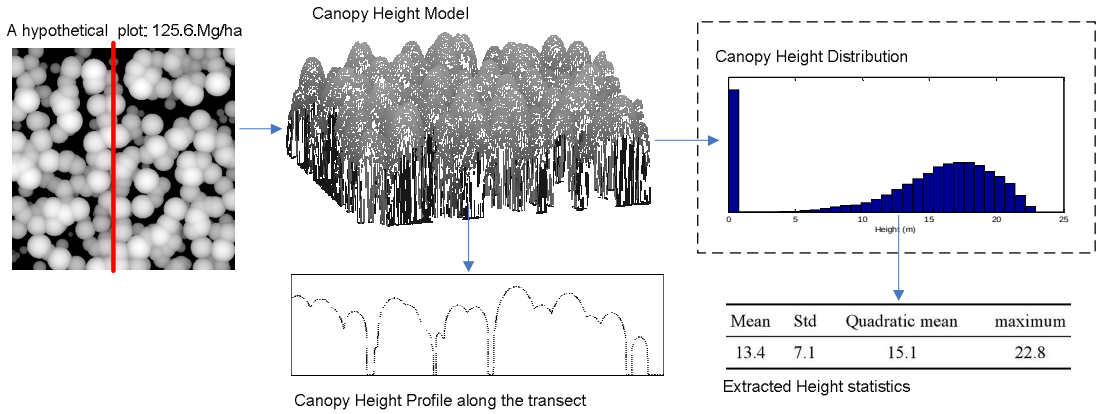


Figure 4.3. A hypothetical scenario to illustrate the flow of diminishing information contents: A simulated plot with hemi-ellipsoid crowns as well as an overlaid transect to mimic a PALS profiling flight line (the left), the Canopy Height Model in a 3-D perspective (the middle upper), the 1-D Canopy height Profile corresponding to the PALS transect (the middle lower), the Canopy Height Distribution derived from the CHM (the right upper), and several height statistics extracted from the CHD.

4.4.3 LiDAR-derived Fine-resolution AGBM Map as Ground-reference Data

Tree Location, Height and Crown Diameter from CHM

The raw LiDAR data were first processed to create a canopy height model (CHM), as outlined in Figure 4.2. CHM represents heights of the top of canopy surfaces relative to grounds. It can be computed as the difference between Digital Surface Model (DSM) and DEM. We created a DSM by first selecting the highest LiDAR point within each .5 x .5 m cell that contains at least one laser hit and then interpolating the selected points into a regular grid of 0.5 x 0.5 m by triangulation. Next, pixel-wise subtraction of

DEM from DSM gave the CHM. Similar procedures to create CHM were frequently reported in the literature where a different cell size or interpolation algorithm may be used.

Subsequently, we applied the TreeVaw software on the CHM to identify individual trees (Popescu and Wynne, 2004). The software implemented a local maximum filtering with a variable circular window. The window size is determined locally and adaptively according to the CHM height at the window center, assuming that a taller tree has a wider crown. The individual tree parameters derived from CHM by the algorithm, i.e., tree location, tree height and crown width, were recorded for each identified tree.

Individual Tree Crown Base Height (CBH) from LiDAR Multiband Height-bin Data

Height-bin data were generated by first discretizing the 3D spatial domain into 3D array of voxels with prescribed horizontal and vertical (height-bin) resolutions and then recording for each voxel the count of laser hits that fall within that voxel. Unlike CHM, height-bin product makes use of the full set of laser hits, and has been proven to be useful for analyzing 3D forest structure. Moreover, height-bin data can be analyzed as mulit-band images where each horizontal slice at a given height bin is treated as a single-band image. In this investigation, a voxel dimension of 0.5 x 0.5 x 1.0 m was used with up to 31 height bins starting from 0.0 m.

CBH was estimated at tree levels by referring to the height-bin data. Specifically, a vertical profile of laser hits per bin was created for each TreeVaw-identified tree by vertically cutting out of the height-bin data a cylinder that is centered at the tree location and that has a diameter equal to crown with. We calculated CBH as the height at which an abrupt drop in the vertical profile occurs. Interested readers are referred to Popescu and Zhao (2008) for an elaboration on the generation of height-bin data and the algorithm for estimating CBH.

Mapping AGBM at Sub-tree Levels

DBH is a widely used proxy for biomass through allometric equations. LiDARs, however, do not directly measure DBH. Therefore, we incorporated the aforementioned LiDAR-derived tree dimension variables, i.e., tree height, crown width and CBH into linear models to predict DBH, and two species-specific linear models were developed by relating the field-measured DBH and the associated LiDAR-derived tree structural variables of 117 matched trees including 94 pines and 27 deciduous trees. The resulting models have a R^2 of 0.79 for pines, and 0.74 for the deciduous.

The classified QuickBird image was used to determine the species of LiDAR-derived trees, namely, pine or deciduous. The diameter-based general allometric equations for pine and mixed hardwood as compiled in Jenkins et al. (2003) were applied to the LiDAR-based estimates of DBH to compute above-ground component biomass. The estimated component biomass of each LiDAR-derived tree was then used to generate a spatially-explicit biomass map at a resolution of .5 m. The generation process proceeded tree by tree: The bole biomass was assigned to the pixel at the tree location and the foliage biomass was distributed uniformly over the pixels covered by the crown. The resultant map is shown in Figure.4.4 where we can clearly see patterns of trees and spatial distribution of AGBM.

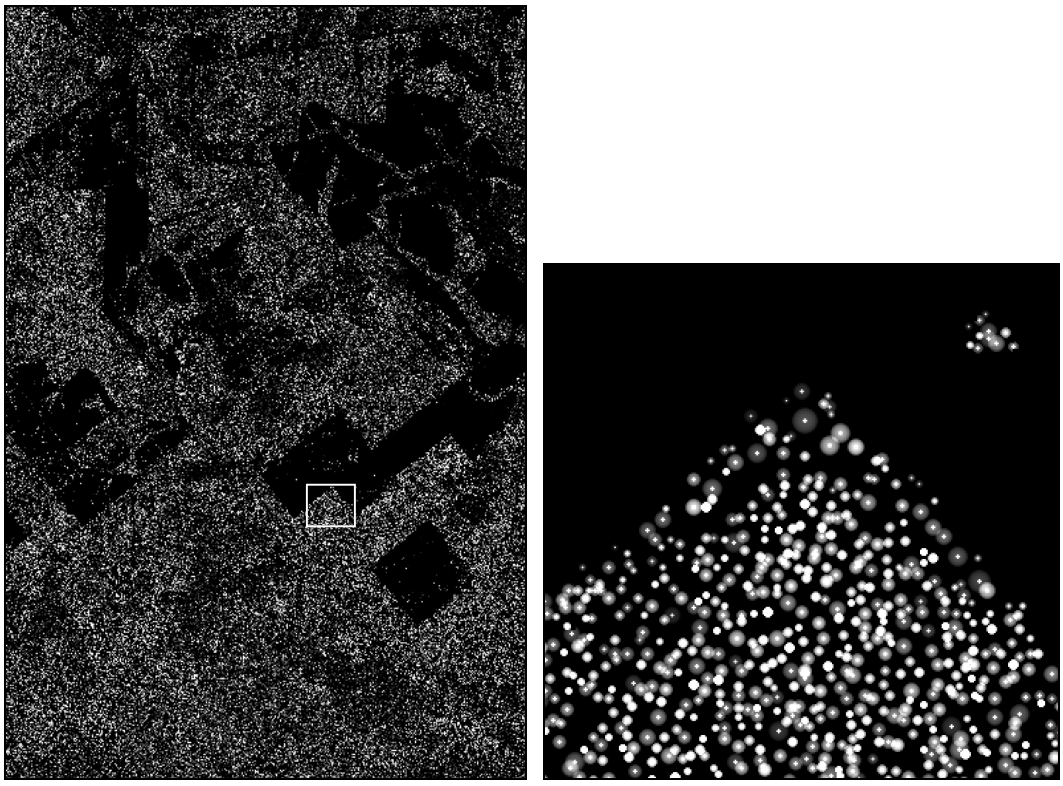


Figure 4.4. The detailed biomass map with a spatial resolution of 0.5 x 0.5 m that was derived from LiDAR scanner data (left) and a close-up of the subset highlighted by the white rectangle (right). The map uses a gray scheme such that brighter pixels indicate higher biomass.

4.4.4 ModelingCanopy Height Distributions (CHD)

This sub-section is dedicated to developing a simple CHD model to relate plot-level CHDs and the distributions of DBH with the aid of tree allometric relationships. Rather than to be theoretically rigorous, the modeling purpose is to provide theoretical justification for the scale-invariant biomass models that will be proposed next, considering that AGBM is dictated by DBH and LiDAR provides measurements on CHDs. Before proceeding, it is helpful to first clarify variables and notations:

h "Canopy height" which is the height of the top surfaces that are visible from above, and refers to both open grounds ($h = 0$) and canopy surfaces ($h > 0$).

$p(h)$ " Canopy height distribution" (CHD), i.e., the probability distribution of h on a unit forested area, which refers to the probability of observing the portion of the unit area that has a height of h, and consists of two terms, i.e., one for the open ground and another for the canopy (Figure.4.3).

c "Canopy cover" which defines the fraction of a unit area obscured by vertically projected crowns when assuming that crowns are opaque.

n "Stem density" which is the number of trees per unit area.

α "Crown projection area" (CPA) which refers to the vertically projected area of a single tree crown onto the ground.

$p_\alpha(\alpha)$ "Distribution of CPA" which specifies a distribution of CPA for trees on a unit area.

H "Tree Height".

D "DBH" of a tree.

h_{cbh} "Crown base height".

$p_D(D)$ "Distribution of DBH" over a unit area.

$p_H(H)$ "Distribution of tree height" for all trees over a unit area.

$p_{tr}(h \mid H, \alpha, h_{cbh})$ "Per-tree crown height distribution" conditioned on tree dimension parameters H, α, and h_{cbh}. The distribution is defined for a tree concerning the canopy height over its CPA, and it is determined by the geometry of the tree.

In the above, the definition of $p(h)$ is made more compatible with that of LiDAR CHM, but $p(h)$ is different from the height distribution of raw LiDAR hits that may be reflected from inside crowns. Of note is that α, D, H, and h_{cbh} are defined for single trees while n, h, and c for an unit area. Other relevant variables will be introduced when necessary.

The variables and distributions defined above are related to each other, e.g., through allometric equations. First, the canopy cover c of a unit forested area is given by the Boolean model (Li and Strahler, 1992),

$$c = 1 - \exp(-n \cdot E[\alpha]) \tag{4.1}$$

where $E[\alpha] = \int \alpha\, p_\alpha(\alpha) d\alpha$ is the mean per-tree CPA. For simplifying illustrations but without losses of generality, we assumed the first-order approximation of Eq. 4.1, i.e., $c \approx n \cdot E[\alpha]$, which corresponds to the case of non-overlapping crowns.

Next, given *c*, the CHD is composed of two terms,

$$p(h) = (1-c) \cdot \delta(h) + c \cdot p_c(h) \tag{4.2}$$

where $\delta(h)$ is the Dirac function that is infinity at $h = 0$ and is 0 otherwise. The first term $(1-c) \cdot \delta(h)$ means that the open ground is seen with a probability of (1-*c*), i.e., the proportion of open ground. The second term $c \cdot p_c(h)$ means that when viewed vertically downward, crowns can be seen with a probability of *c*, i.e., the proportion of the aggregated CPAs. $p_c(h)$ is the seen CHD over the canopy only ($h > 0$), and it can be elucidated as a CPA-weighted sum of crown height distributions. If the stem density *n* is relatively large, $p_c(h)$ is approximated by,

$$p_c(h) \approx \int p_{tr}(h \mid D) \frac{\alpha(D)}{E[\alpha]} p_D(D) dD \tag{4.3}$$

where $p_{tr}(h \mid D)$ is the per-tree crown height distribution conditional on D, and $\alpha(D) = E[\alpha \mid D]$ is the mean per-tree CPA given *D*.

To model $p_{tr}(h \mid D)$ of Eq. 4.3, let us first address $p_{tr}(h \mid H, \alpha, h_{cbh})$, which can be explicitly modeled if the crown geometry is unknown. Also, note that CPA α can be explicitly computed from crown diameters if crown shapes take some known geometry primitives, e.g., ellipsoid and cone. For some crown geometries, it holds that $p_{tr}(h \mid H, \alpha, h_{cbh}) = p_{tr}(h \mid H, h_{cbh})$. Meanwhile, allometric relationships, whether deterministic or probabilistic, can be used to derive the distributions of H, α and h_{cbh} conditioned on *D*, i.e., $p_H(H \mid D)$, $p_\alpha(\alpha \mid D)$ and $p_{h_{cbh}}(h_{cbh} \mid D)$, (Tewari and Gadow, 1999; Li et al., 2002). Let us further assume the independence of the conditional

distributions of H, α and h_{cbh} given D. As a result, we obtain the per-tree crown height distribution conditioned on D,

$$p_{tr}(h \mid D) = \int p_{tr}(h \mid H, \alpha, h_c) p_H(H \mid D) p_\alpha(\alpha \mid D) p_{h_c}(h_c \mid D) dH \cdot d\alpha \cdot dh_c \quad (4.4)$$

which generally depends on tree species. Nevertheless, if there are multiple species, e.g., $p_{tr}(h \mid D, i)$, $i = 1, \ldots, m$, they could be assimilated into an overall distribution that is unconditioned on species by assuming that the occurrence probabilities of each species, r_i, could be elicited using some prior knowledge $p_r(r_1, r_2 ..., r_m)$ where $\sum r_i = 1$. Often, it is convenient to model $p_r(r_1, r_2 ..., r_m)$ with a multinomial distribution. Marginalizing over $r_1, r_2 ..., r_m$ yields

$$p_{tr}(h \mid D) = \int (\sum_{i=1}^{m} p_{tr}(h \mid D, i) \cdot r_i) \cdot p_r(r_1, r_2, ..., r_m) dr_1 dr_2 \cdots dr_m \quad (4.5)$$

To finalize the model description, we substituted Eqs .4.1 and 4.3 into Eq. 4.2 and obtained an expression for CHD,

$$p(h) = (1 - n \cdot E[\alpha]) \cdot \delta(h) + n \cdot \int p_{tr}(h \mid D) \alpha(D) p_D(D) dD \quad (4.6)$$

which establishes a link between CHD ($p(h)$) and the distribution of DBH ($p_D(D)$) through the known crown geometry and allometric relationships. On the other hand, the AGBM over the unit forested area, M, is readily calculated from $p_D(D)$ and n with the recourse to the DBH-based biomass allometric equation $B(D)$,

$$M = n \int B(D) \cdot p_D(D) dD \quad (4.7)$$

4.4.5 A Scale-invariant Linear Functional Model for Biomass

It becomes clear from Eqs. 4.6 and 4.7 that CHD $p(h)$ is related with biomass M through $n \cdot p_D(D)$. Once an estimate of CHD $\hat{p}(h)$ is extracted from LiDAR data or LiDAR-derived products such as CHMs, the inversion of Eq. 4.6, if possible at all, provides knowledge about $n \cdot p_D(D)$ that then can be incorporated into Eq. 4.7 to infer

M . As per this logic, we proposed the use of CHD $p(h)$ as a predictor for estimating AGBM. Also, denote the AGBM model by $M = f[p(h)]$, which is a functional because the argument $p(h)$ is a function of h by itself. More interestingly, we gleaned an extra benefit that the predictor, $p(h)$, is linearly scalable, which means, for example, that if $p_1(h)$ and $p_2(h)$ are CHDs defined for two contiguous regions with areas of a and b, respectively, the CHD for the combined region, $p_{12}(h)$, is an area-weighted linear combination of $p_1(h)$ and $p_2(h)$,

$$p_{12}(h) = \frac{a \cdot p_1(h) + b \cdot p_2(h)}{a + b} \tag{4.8}$$

For the AGBM model to be scale-invariant, f must satisfy,

$$f[\frac{a \cdot p_1(h) + b \cdot p_2(h)}{a + b}] = \frac{a \cdot f[p_1(h)] + b \cdot f[p_2(h)]}{a + b} \tag{4.9}$$

where a, b, $p_1(h)$ and $p_2(h)$ assume certain arbitrariness; therefore, $f[\cdot]$ should be a linear functional of $p(h)$. A natural choice for such an f is $f[p(h)] =< p(h), K(h) >$ where $<\cdot,\cdot>$ denotes an inner product in function spaces and $K(h)$ is a fixed function that needs to be inferred, e.g., from training data. For simplicity, we chose the most common inner product and proposed the following scale-invariant AGBM model,

$$M = f[p(h)] =< p(h), K(h) >= \int_{h \geq 0} p(h) \cdot K(h) dh \tag{4.10}$$

In practice, $p(h)$ is estimated or extracted from LiDAR measurements, and its function values are available only at a discrete set of points, i.e., $\{p(h_i)\}_{i=1}^{k}$, $h_i = (i-1) \cdot \Delta h$. Accordingly, from Eq. 4.10, f can be approximated by,

$$M \approx f[\{p(h_i)\}_{i=1}^{k}] = \sum_{i=1}^{k} p(h_i) K(h_i) \Delta h \tag{4.11}$$

where Δh is a height step (height-bin) used to sample $p(h)$. In essence, Eq. 4.11 is a linear model that has $\{p(h_i)\}_{i=1}^{k}$ as predictors with the corresponding coefficients being $K(h_i) \cdot \Delta h$.

4.4.6 A Scale-invariant Nonlinear Functional Model for Biomass

Because a quantile function is mathematically equivalent to its associated probability density and they can be converted to each other without any loss of information, instead of using the density CHD $p(h)$, we also attempted to use canopy height quantile function (CHQ) as an alternative predictor for scale-invariant biomass prediction. However, one undesirable feature of CHQs is their nonlinear scaling property, by which we mean that the overall CHQ over a forested region cannot be generally aggregated as a linear combination of the multiple CHQs over the sub-regions that partition the region. This scale-dependence indicates that linearity and scale-invariance are unlikely to be achieved at the same time for a model using CHQs as predictors; hence, a nonlinear model is needed when using CHQs as predictors. Considering the equivalence of the CHQ to the CHD, we derived a nonlinear version of biomass model by first transforming CHQ into CHD and then substituting the transformed CHQ into the aforementioned linear model in Eq. 4.10. For this purpose, let $F(h)$ be the cumulative CHD, i.e., $dF(h)/dh = p(h)$. The inversion of $F(h)$ is the CHQ $Q(q)$, i.e., $Q(q) = F^{-1}(q)$ where $0 \le q \le 1$. Using this CHD-CHQ relationship as well as the integration-by-parts theorem, Eq. 4.10 can be re-written as,

$$M = K(h_{\max}) - \int_0^1 q \cdot \frac{dK(Q(q))}{dq} dq \tag{4.12}$$

where $h_{\max} = \min(h; F(h) = 1)$ is the maximum canopy height observed over the analysis unit, and $K(\cdot)$ is the same function as introduced in Eq. 4.10. Eq. 4.12 is a biomass model using the CHQ $Q(q)$ as predictor, and $K(\cdot)$ is fixed but need to be estimated either by resorting to training data or theoretical calculation.

As with $p(h)$, $Q(q)$ are practically observed only at a discrete set of q's, i.e., $\{h_i = Q(q_i)\}_{i=1}^{k+1}$ where $q_i = (i-1)/k$. With this discrete set of values $Q(q_i)$, we arrived at the following approximation of the nonlinear biomass model,

$$M \approx \sum_{i=1}^{k} K(h_{i+1})/k \tag{4.13}$$

where h_{i+1} is the (i+1)-th (k+1)-quantiles of CHD. This model is believed to be nonlinear with respect to h_i due to the potential nonlinearity of $K(h)$, but it essentially is equivalent to the linear one in Eq. 4.13,

4.4.7 Derivation of the Scale-invariant Biomass Model from the CHD Model

The two biomass models proposed above can be directly uncovered from the CHD model formulated in Eqs. 4.1-4.7. Because the two models are equivalent except for the use of different predictors, the following will focus only on the derivation of the linear model of Eq. 14. By comparing Eq. 4.2 with Eq. 4.6, we isolated the canopy part of CHD, i.e., the second term of $p(h)$, as given below,

$$c \cdot p_c(h) = \int p_{tr}(h \mid D) \cdot \alpha(D) \cdot n \cdot p_D(D) dD \tag{4.14}$$

which essentially is a Fredholm integral equation of the first kind with an aim to solve out $n \cdot p_D(D)$ given $c \cdot p_c(h)$ and $p_{tr}(h \mid D) \cdot \alpha(D)$ (Aster et al., 2004). In practice, $c \cdot p_c(h)$ may be estimated from LiDAR CHM by discarding the ground component from $p(h)$, and $p_{tr}(h \mid D) \cdot \alpha(D)$ may be derived from empirical relationships such as height-diameter equations. Once $n \cdot p_D(D)$ is solved from Eq. 4.14, it can be incorporated into Eq. 4.7 with a known allometric equation $B(D)$ to compute AGBM.

To derive the linear functional model, we present a symbolic scheme for solving Eq. 4.14, although this is of little practical value due to its numerical unstability of an ill-posed problem. First, denote $p_{\text{sec}}(h) = c \cdot p_c(h)$ for notional convenience; then discretize the integral equation using $\{h_i = (i-1) \cdot \Delta h\}_{i=1}^{k}$ for h and $\{D_j = (j-1) \cdot \Delta D\}_{j=1}^{l}$ for D. As a result, Eq. 4.14 reduces to,

$$\mathbf{p}_{\text{sec},(k \times 1)} = \mathbf{A}_{(k \times l)} \cdot \mathbf{p}_{n,D,(k \times l)} \tag{4.15}$$

where $\mathbf{p}_{\text{sec},(k\times 1)} = [p_{\text{sec}}(h_1),...,p_{\text{sec}}(h_k)]^T$, $\mathbf{p}_{n,D,(l\times 1)} = [n\cdot p_D(D_1),...,n\cdot p_D(D_l)]^T$, and $\mathbf{A}_{(k\times l)}$ is a $k\times l$ matrix with its (i, j) element being $p_{tr}(h_i \mid D_j)\cdot\alpha(D_j)\cdot\Delta D$. Applying $\mathbf{A}^{-1}_{(k\times l)}$, i.e., the generalized inverse of $\mathbf{A}_{(k\times l)}$, to Eq. 4.15, $\mathbf{p}_{n,D,(l\times 1)}$ is obtained as,

$$n\cdot p_D(D_j) = \sum_{i=1}^{k} a_{ji}^{-1}\cdot p_{\text{sec}}(h_i) \tag{4.16}$$

where a_{ji}^{-1} is the (j, i) element of $\mathbf{A}^{-1}_{(k\times l)}$. Combining Eq.7 and Eq. 4.16, AGBM can be estimated by,

$$M \approx \sum_{i=1}^{k} p(h_i)\cdot\sum_{j=1}^{l} a_{ji}^{-1}\cdot B(D_j)\cdot\Delta D \tag{4.17}$$

which is a linear model of the same form as Eq. 4.13. A comparison between Eq. 13 and Eq. 17 reveals the following relationship,

$$K(h_i) = \sum_{j=1}^{l} a_{ji}^{-1}\cdot B(D_j)\cdot\frac{\Delta D}{\Delta h} \tag{4.18}$$

4.4.8 K-function

The newly introduced function $K(h)$ acts like a weighting function in the integral of Eq. 1, and represents the contribution to per-area biomass by the portion of canopies that has a height falling within/around h. However, $K(h)$ cannot be an arbitrary function. It is beneficial to impose certain constraints on it with recourse to intuitive prior knowledge. First of all, we supposed that the AGBM becomes zero if no tree is present ($p(h) = \delta(h)$), i.e., $M(\delta(h)) = \int \delta(h)K(h)dh = 0$; this indicates $K(0) = 0$. Moreover, we assumed that $K(h)$ is non-decreasing with respect to h, considering that the larger the percentage of tall trees, the higher AGBM.

Numerically, $K(h)$ can be solved out by inverting the following Fredholm integral equation of the first kind,

$$B(D) = \int_{h\geq 0} K(h)\cdot p_{tr}(h \mid D)\alpha(D)dh \tag{4.19}$$

which is derived by substituting $p(h)$ of Eq. 4.6 into Eq. 4.10 and then relating Eq. 4.10 with Eq. 4.7. The inversion of Eq. 4.19 for $K(h)$ with known allometic relationships including $B(D)$ and $p_{tr}(h \mid D)\alpha(D)$ is an ill-posed problem that usually requires regularization schemes for numerical solutions. A symbolic solution to Eq. 4.19 for $K(h)$ is the same as in Eq. 4.18.

In this investigation, we made no attempt to infer $K(h)$ by directly solving the integral equation or using Eq. 4.18 because this involves explicit modeling of $p_{tr}(h \mid D)$. Instead, we turn to a constrained regression technique to infer $K(h)$ through the linear functional model in light of training data sets, as detailed below.

4.4.9 Synthesized Training and Testing Datasets

Synthesized data were simulated to test and train the proposed models. In the datasets, the predictor variables extracted from the CHM are the discretized CHDs $\{p(h_i)\}_{i=1}^{k}$ and CHQs $\{h_i = Q(q_i)\}_{i=1}^{k+1}$ for the linear and nonlinear models, respectively. The response variables (i.e., dependent variables) are the associated AGBM obtained by aggregating the detailed biomass map on the coincident plots. In the process of synthesis, we investigated two plot shapes, both circular and square. For each shape, we used 10 different plot sizes that are $(10i)^2$ m^2 with *i* ranging from 1 to 10, thus resulting in a total of 20 shape-by-size combinations. Furthermore, for each combination, a total of 2000 simulated plots (data points) were sampled systematically over the study area, thus yielding 40000 data points.

4.4.10 The Constrained Least-squares Method for Model Inference

For simplicity, the discretized linear model of Eq. 4.13 is used to infer the coefficients $K(h_i)$ (i.e., the discretized version of $K(h)$), although other complicated or possibly more effective quadrature methods may be referred to for the discretization of the integral equation in Eq. 13 (Aster et al., 2004). However, even with Eq. 4.13,

common regression procedures fail to work out partly due to the constraints imposed on $K(h)$. Alternatively, we chose a constrained least-square procedure, as implemented in Matlab (*lsqlin* function, Mathworks, Inc.), to iteratively search the optimal values of $K(h_i)$ in terms of minimizing the mean square error with regard to the training data. The constraints took both inequality and equality forms such as $K(h_i) \geq K(h_{i-1})$ for $i > 1$ and $K(h_0 = 0) = 0$.

4.5 Results

4.5.1 Evaluation of the Linear Functional Model

In the synthesized data, we first used a 0.25 m height bin (sampling interval) to discretize the CHDs up to 35 m, therefore, resulting in a predictor vector of CHD that contains 140 scalar values. For a given group of 2000 plots of a given size and shape, we repeatedly split the group randomly into 100 datasets of 100 training and 1900 test plots to assess the consistency of the models across different training/test datasets.

The regression results are presented in Table 4.2, which includes the mean, minimum and maximum of coefficients of determination (R^2) as well as the mean root mean square errors (RMSE) over each group of 100 runs. It is observed that for all plot sizes, plot shapes had no significant impacts on model inferences in terms of both R^2 and RMSE (all p-values << 0.01). Overall, R^2 values increase with an associated decrease in RMSE as the plot size varies from 0.01 ha to 1.0 ha. For example, in a random trial using the 0.01-ha square-plot data, R^2 for training and testing are 0.85 and 0.83, respectively, with the associated RMSEs of 30.3 and 31.2, respectively, (Figure 4.5); and in the case of 1-ha square-plot data, R^2 for training and testing are both about 0.94, with a RMSE of 14.1 and 14.4, respectively (Figure 4.5). In all 2000 cases we evaluated (2 plot shapes by 10 plot sizes by 100 random runs), performances of the linear functional model in training are consistent with those in testing as indicated by a good agreement of R^2 or RMSE (all p-values << 0.01) between training and testing (Table 4.2). Such

consistence partially validates the proposed linear model, and also shows its good generalization ability.

Table 4.2. The coefficients of determination (R^2) and RMSEs of the linear model (140 predictors with a height bin of 0.25 m) trained and tested across a series of plot sizes for both square and circular shapes. Note that, in each case, R^2 is reported as the mean, minimum and maximum over 100 random runs, and RMSE is reported as the mean over the 100 runs.

Plot size	Square Plots				Circular Plots			
(ha)	Training		Testing		Training		Testing	
	R^2 (min,max)	rmse	R^2 (min,max)	rmse	R^2 (min,max)	rmse	R^2 (min,max)	rmse
.01	.832 (.804, .855)	33.3	.825 (.813, .830)	34.2	.831 (.805, .854)	33.7	.823 (.815, .828)	34.6
.04	.892 (.877, .908)	24.3	.889 (.885, .892)	24.7	.892 (.874, .909)	24.3	.887 (.882, .891)	25.0
.09	.912 (.897, .922)	21.0	.910 (.907, .913)	21.3	.912 (.897, .928)	19.9	.911 (.907, .913)	21.3
.16	.921 (.910, .933)	19.2	.919 (.917, .921)	19.5	.921 (.909, .936)	19.1	.919 (.916, .921)	19.6
.25	.926 (.914, .934)	18.0	.925 (.923, .927)	18.2	.925 (.914, .934)	18.1	.924 (.923, .927)	18.3
.36	.931 (.919, .943)	16.9	.929 (.927, .931)	17.2	.932 (.921, .945)	16.9	.929 (.927, .930)	17.3
.49	.933 (.923, .943)	16.2	.932 (.930, .934)	16.4	.932 (.920, .943)	16.3	.932 (.930, .934)	16.5
.66	.937 (.925, .948)	15.4	.934 (.933, .936)	15.7	.935 (.924, .946)	15.6	.934 (.932, .936)	15.8
.81	.937 (.928, .948)	15.0	.937 (.934, .938)	15.1	.937 (.927, .947)	15.0	.936 (.935, .938)	15.2
1	.940 (.928, .952)	14.3	.938 (.936, .940)	14.6	.938 (.928, .946)	14.6	.938 (.937, .940)	14.7

Four examples of the regressed $K(h)$ function, as obtained from the series of optimized coefficients of the linear model, are plotted in Figure 4.6 where each of the four curves is the average over the estimates of 100 random runs. Across all the 20 combination of factors considered, i.e., two plot shapes by 10 plot sizes, there is no significant difference (all p-values << 0.01) between any two average $K(h)$ curves as depicted in Figure 4.6, although minor differences may be observed , especially at the end of curves or between the estimated curves from individual runs even for the same plot size.

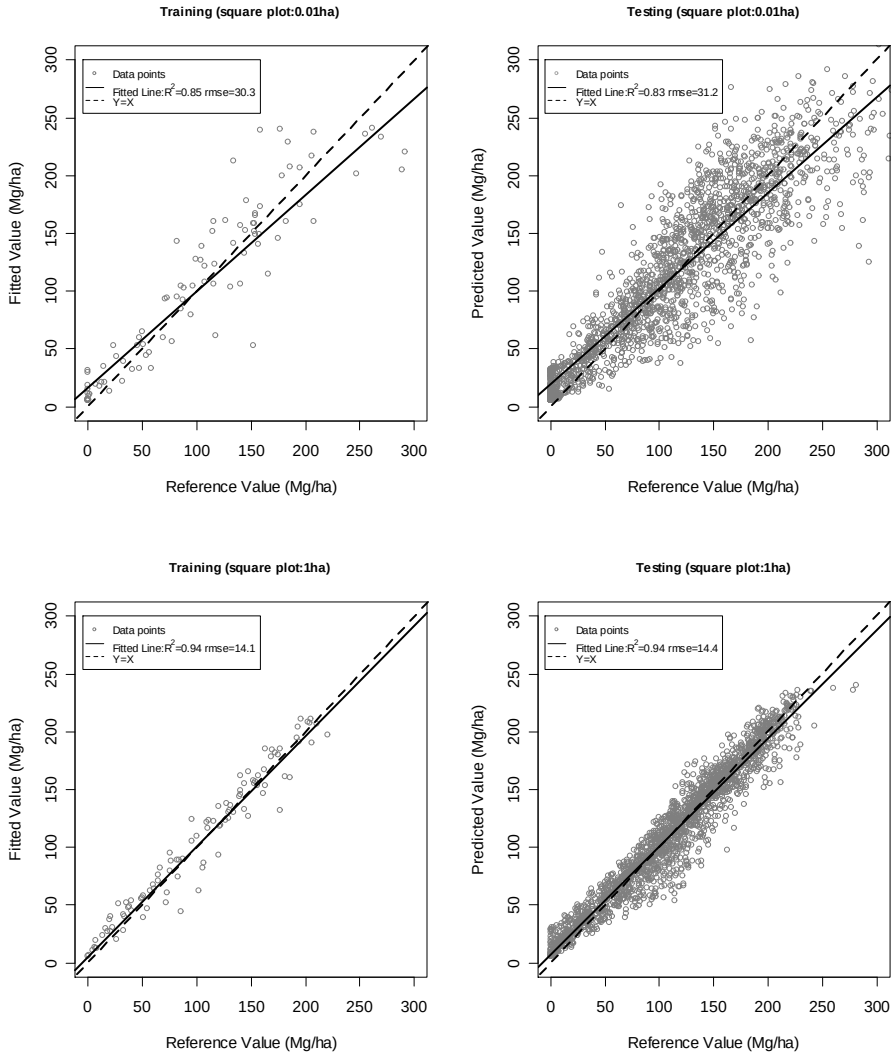


Figure 4.5. Scatterplots of predicted vs. reference biomass for the training (the left two) and the testing (the right two), respectively, and also over the plot sizes of 0.01 ha (the upper two) and 1 ha (the lower two), respectively. The model is the linear model with 140 predictors (height bin: 0.25 m).

4.5.2 Evaluation of the Nonlinear Functional Model

It is not an easy endeavor to estimate $K(h)$ directly from fitting the nonlinear model in Eq. 22. Therefore, we applied the $K(h)$ estimated for the linear model to the nonlinear model to show the equivalence of the two models. The predictors in the

nonlinear model are heights associated with CHQs which we discretized at $q_i = (i-0.5)/100\%, i = 1,2,...,100$, as well as 99.6%, 99.7%, 99.8% and 99.9%. Because the estimated values of $K(h)$ are available only at discrete heights, we interpolated these values to calculate $K(h)$ at an arbitrary height for the nonlinear model. To assess the equivalence of the nonlinear model to the linear model, we compared the biomass prediction of the two models in the form of scatterplots where, for brevity, only two cases are depicted: one for the 0.01 ha square plot and another for the 1.0 ha square plot (Figure 4.7). In both cases, the two models matched each other exactly for all but several testing data points. Most of the exceptions are due to truncation errors in the CHDs of the linear model that have been discreizted only up to 35 m. As an example, Figure 4.8 illustrates a truncated CHD that corresponds to the outlier plot circled in Figure 4.7(a).

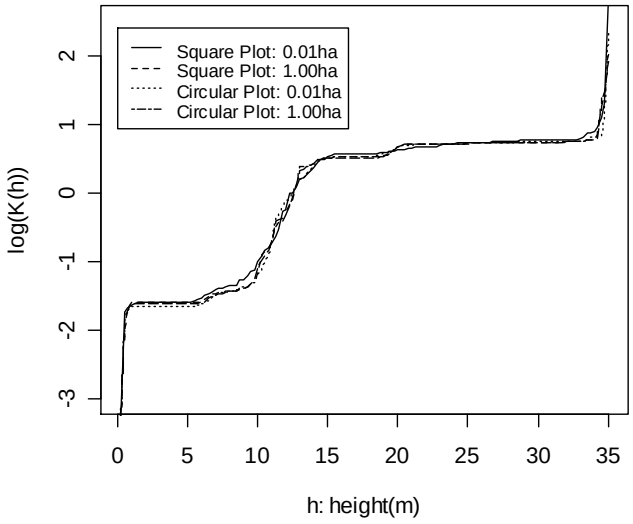
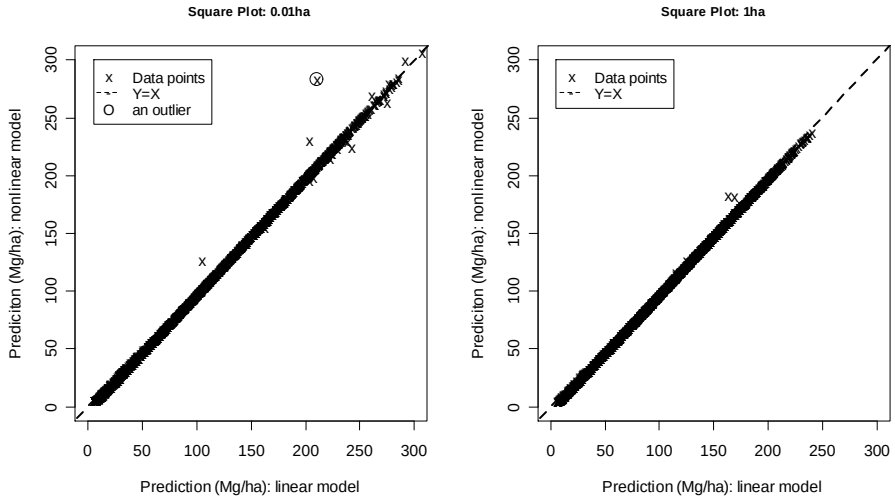


Figure 4.6. K-functions: The estimated coefficients of the linear model are plotted as curves with respect to h. Each curve is the average over the respective 100 runs.



(a) (b)

Figure 4.7. The scatterplots of prediction by the nonlinear model versus those by the linear model for (a) the 0.01 ha plot-level data and (b) the 1ha plot-level data.

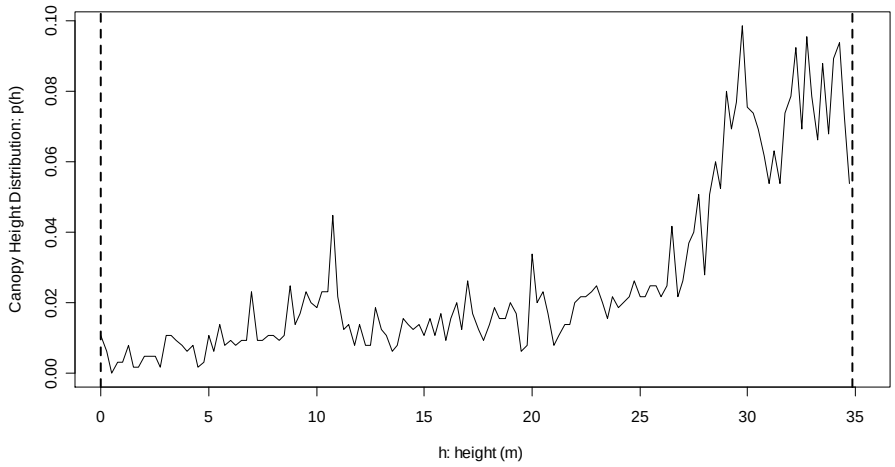


Figure 4.8. An example of truncated canopy distributions where portion of canopy above 35 m are not taken into account due to the discretization scheme: This example corresponds to the outlier case circled in Figure 4.5(a).

4.5.3 Model Analysis

To provide practical guides, we examined the effects of the length of height bin (a step in height to sample CHD) and the size of training dataset on performances of the linear model. The choice of a sampling interval or height bin not only affects the characterization of discretized CHD, but also determines the dimension of CHD predictor with a larger height bin corresponding to lower dimension in the same way as that a spectral signature is characterized by different spectral resolution (Figure 4.9). With a training data set size of 100 and a test set size of 1900 for all plot sizes, we evaluated the linear model by using a series of height bins ranging from 0.25 m to 20 m that corresponds to a predictor dimension ranging from 140 to 2. Typical results of R^2 in training and testing are demonstrated for 0.01-ha and 1.0-ha square plots, respectively (Figure 4.10). It is noted that the model performances showed no degrading trend until the height bin increases above around 6.0 m, and that the R^2 dropped significantly when using a height bin larger than 12.0 m. This observation suggests that it is practical and effective to choose a relatively large height bin below a critical value with no great loss of model performances. The critical value may vary with respect to specific applications, e.g., around 5 to 6 m in this investigation.

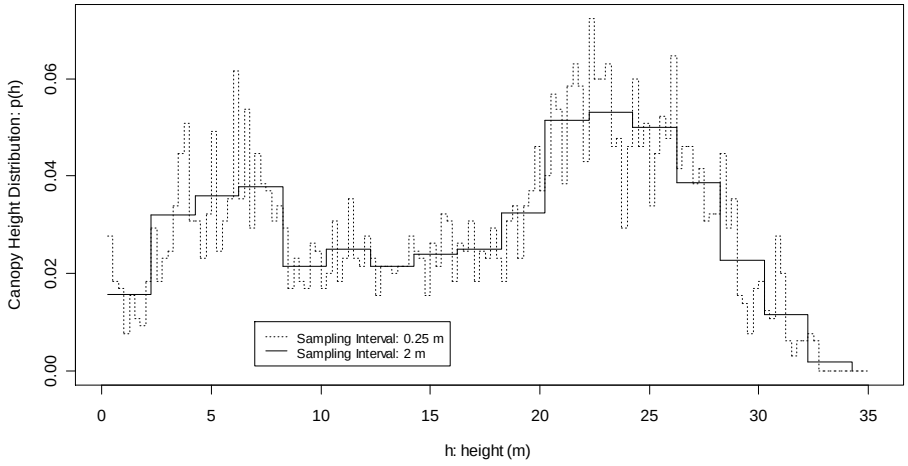


Figure 4.9. An example showing that the same canopy height distribution is discretized with two different height bins. A larger height bin leads to a coarser-resolution canopy height distribution.

In addition to the height bin, we varied the size of training data from 10 to 500 with an increment of 10 to evaluate the effects of training sizes on model performances, and the models trained with these different training data sizes were tested on the same independent set of 1500 data points. Meanwhile, we also varied the CHD dimension from 7 to 140 that corresponds to a height bin ranging from 5 m to 0.25 m in disretizing the CHDs. In this analysis, all evaluations were based on 0.49-ha square-plot data because plot shapes and sizes have no or little influences on the model inference. To avoid plethora of figures, we reported the changes of R^2 with training data sizes only for two extreme models, i.e., one with a predictor dimension of 140 (height bin: 0.25 m) and another with a predictor dimension of 7 (height bin: 5 m) as shown in Figure 4.11. It is found that, for all the models considered, the addition of new data into training sets with a size of above 70 contributed to no significant gain in terms of testing R^2, although the R^2 for training may fluctuate a little bit which suggests a possibility of slightly overfitting in some cases. The result indicates that, in our cases, a training size of 50 is typically enough for the purposes of fitting the model and making prediction. However, caution should be exercised that in practice, training data should be collected over a wide range of conditions so that no extrapolation is risked.

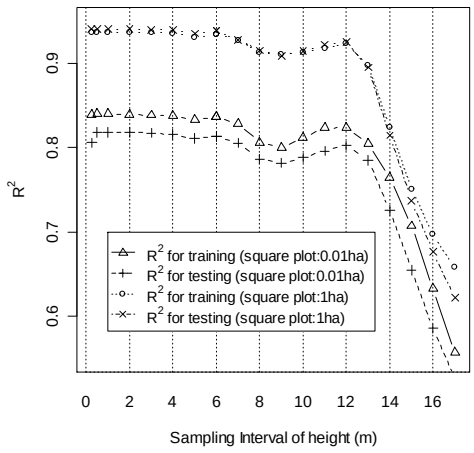


Figure 4.10. R^2 vs. Sampling interval of height (height bin): Changes in R^2 for training and testing as a function of the sampling interval of height or height bin used to discretize the canopy height distributions.

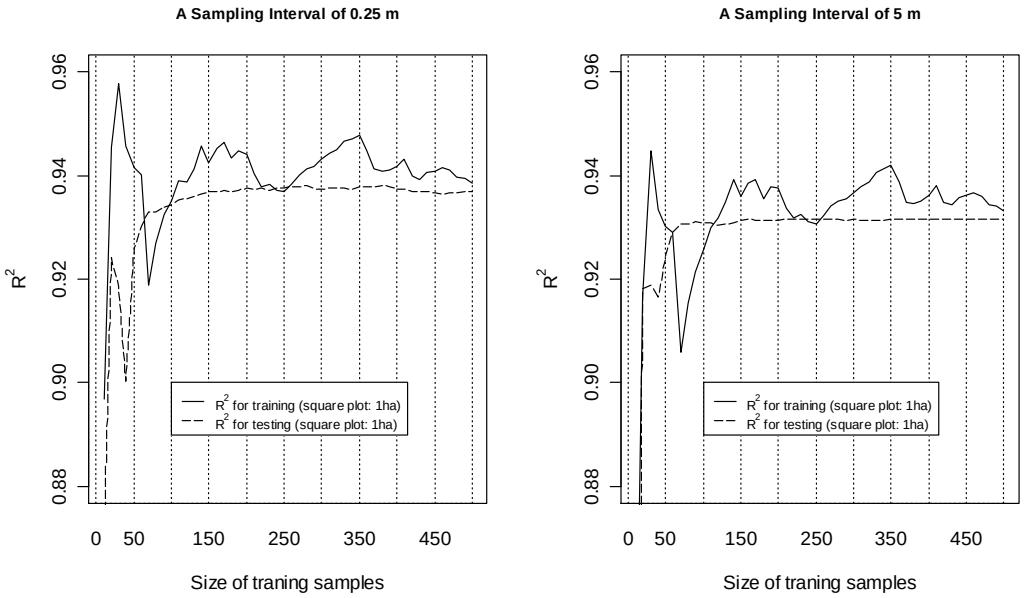


Figure 4.11. The effects of training sample sizes on the R^2 for two different height bins: 0.25m (the left) and 5.0 m (the right). The dataset used is the 1 ha plot-level data.

4.5.4 Scaling-up for Overall Biomass Prediction

The linear functional model was assessed in an extreme case by treating the whole study area as one large analysis unit where a single CHD with a dimension of 140 was extracted as an overall predictor to predict the overall biomass. Only the models trained on square-plot data were examined, and there were 100 random runs for each plot size. These overall AGBM estimates are presented in the box-whisker plots of Figure 4.12 where each column summarizes the estimates from 100 runs for a certain plot size. On average, all the models overestimated AGBM in comparison to the reference value of 109.0 Mg/ha.

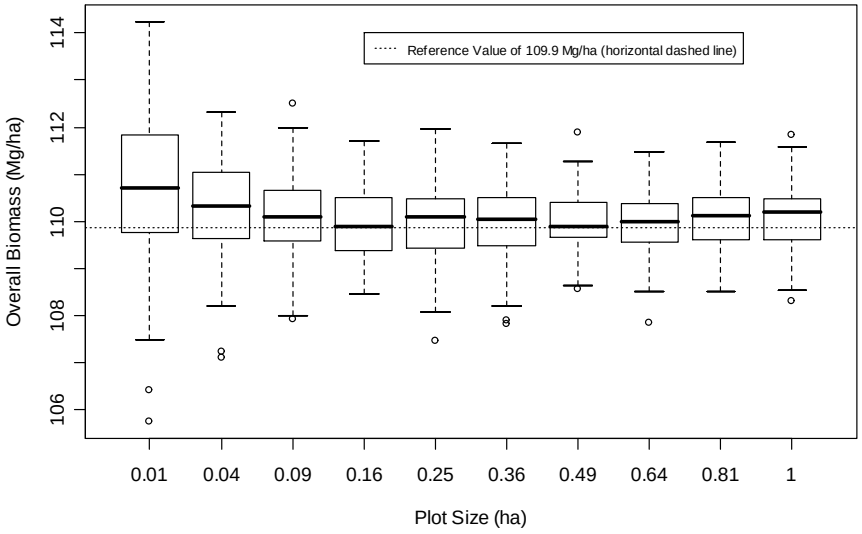


Figure 4.12. The box-whisker plot of the overall biomass estimated by using the single overall CHD of the whole scanner LiDAR CHM. For each plot size, there are 100 models obtained respectively from 100 runs with different random training sets. The 100 models applied to the overall canopy height distribution then produced 100 estimates of the overall biomass.

To assess the effectiveness of profiling laser measurements in predicting overall biomass, we enumerated all the combinations of six PALS transects where, for each given number of transects, i.e., m, there are in total $6 \cdot 5 \cdot (6-m+1)/m!$ possibilities. It is obvious that only one possibility exists when using all the six transects. The profiler-based CHDs were extracted from the pooled PALS transects of each combination. The linear model used is the mean $K(h)$ averaged over the 100 runs for the 1.0 ha square-plot data. It is found that, on average, there is an overestimate for any given number of transects as compared to the reference biomass value obtained by averaging the scanner LiDAR-derived biomass map (Figure 4.13). As also noted, there is a decrease in the variability of estimates when using more PALS transects because PALS provides incomplete observation of the study area only along several transects and the addition of extra transects is expected to lead to better inference about the overall CHD.

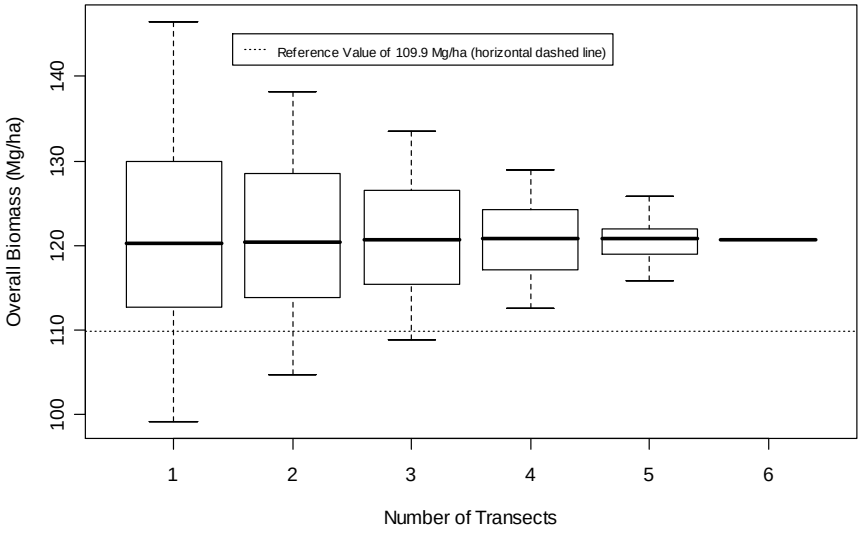


Figure 4.13. The box-whisker plot of the estimated overall biomass using the six PALS profiling transects. Each column represents the estimates obtained from all the possible transect combinations that draw the given number of transects from the six available.

4.6 Discussion

In the development of scale-invariant biomass models, we chose either CHDs or equivalently CHQs as predictors for the use in functional models. The major tenet behind such choices is that a CHD or CHQ, perceived as a function curve, is able to retain as much information as possible, in comparison to the extracted single statistics of CHDs such as mean height, truncated mean height, and quantile-based height (Næsset, 2002; Lim and Treitz, 2004) . Our models, therefore, have potentials of making the most of information available in LiDAR observations (Figure 4.3), and they also have the advantages of being scale-invariant and mathematically justifiable. The functional forms of our models rely on a newly introduced *K* function that has a specific physical meaning. In this study, we referred to a constrained linear regression procedure to infer the *K* function through the linear biomass model. Alternatively, it is recommended that

future studies investigate the usefulness of advanced data mining techniques such as Support Vector Machines and Gaussian Processes to uncover the possibly nonlinear relationships between biomass and CHDs or CHQs (e.g., high-dimensional data due to the discertization of curves) since previous studies reported the successful applications of these techniques for supervised learning with high-dimensional data (Durbha et al., 2007; Zhao et al., in review).

Also, concomitant with the scale-invariance, our models feature shape- and translation invariance, by which we mean that the model outputs for a forested region remain the same irrespectively of the choices of modifiable analysis units that partitions the study region as well as the origins where these units are placed. We suspect, however, that most previously developed models are subject to not only scale-dependence as already recognized by the authors (Næsset, 2002) but also shape- and translation- dependence, although the latter two types of dependence that may also be caused by the nonlinearity of models and the non-scaling property of predictors have not been examined yet. Thus, the scale-invariant models in this study hold great promises for more effective and wider uses in multi-scale forest inventory on variable analysis units that have been obtained by segmenting CHM as reported in van Aardt et al. (2006) and Zhao and Popescu (2007). For example, this study has shown an extreme case of predicting overall biomass from PALS profiling data where the analysis units are individual lines.

Another important aspect about scale or plot size is concerned with our model formulation. By analogy to the formulation of radiative transfer theory in forest canopies (Ross, 1981), the derivation of our conceptual model involves some contradictory assumptions: on the first hand, we expect that the number of trees within a plot is small enough so that not too much overlapping exists to validate the first-order approximation of canopy cover in the Boolean model; on the other hand, we assume that there is a large sample of trees for justifying a continuous distribution. The number of trees largely depends on the plot size that is used to define the problem. Nevertheless, it is advantageous in practice to train a model on large plots, but the choices of plot size or

cell size for prediction purposes make no difference as far as the total biomass in the study area is concerned. However, be aware that creating a biomass map with our approach at an extremely finer resolution, e.g., 0.1 x 0.1 m, makes no physical sense at all. Our models are expected to predict biomass at least above individual tree levels.

The increase in R^2 with plot sizes as noted in Table 4.2 is likely due to plot-edge effects: Because tree stems contains a high percentage of total tree biomass and the methods for extracting predictor and dependent variables count only the components that fall into a plot, discrepancies occur whenever a tree stem falls outside the plot while a large portion of the tree crown falls inside the plot, or vice versa. The edge effect is more pronounced for small plots (Andersen et al.; 2005). However, the edge effect has only minor influence on the model parameter inference (i.e., the *K* function) as has been demonstrated by the statistically indistinguishable mean functions of *K* that were estimated for different plot sizes (Figure.4.6). In practice, it may be desirable to undertake field-work on larger plots while keeping a reasonable cost. If the inventory budget is limited, an alternative is to first construct a computer-based canopy simulator based on the limited ground data and then run the simulator to generate synthesized observations for model development (Nelson, 1997).

Two assumptions, i.e., non-negativity and non-decreasing, have been made about the *K* function. However, more stringent constraints can be imposed on *K*. For example, we can assume that *K* take certain parametric forms. A simple choice is $K(h) = a \cdot h + b$, giving a familiar model,

$$M = \int (a \cdot h + b)\, p(h) dh = a\bar{h} + b \qquad (4.20)$$

where $\bar{h}$ is the mean canopy height. Other rather complicated parametric forms, e.g., $K(h) = a \cdot [\exp(b \cdot h) - 1]$ or $a \cdot [h^b - 1]$, may be explored in further studies. In such cases, the inference of K can boil down to the estimation of its functional parameters only, e.g., a and b in the above examples. In addition, we have presumed that $K(0) = 0$ such that the biomass on a plot empty of trees is zero. This assumption makes our models different from those that produce non-zero biomass for non-tree plots (e.g., zero

canopy height). Patenaude et al. (2004) explained that such initial values at zero LiDAR-recorded height represent the biomass in ground vegetation and litter compartments. In such a sense, the model of this study only considers the biomass present in "standing" trees. However, as to those models of non-zero initial biomass, a possible concern arises that the prediction for zero or small canopy heights may risk anomalous biomass values associated withextrapolation because these low-canopy plots (e.g., clear-cut) are sometimes excluded from the model fitting, e.g., for numerical reasons (Lim and Treitz, 2004). To this end, the model proposed by us has the advantages of being applicable across a wide range of biomass values, and yields exactly zero biomass value for open grounds where no trees are present.

Unlike previous studies that extracted LiDAR-based statistics directly from LiDAR raw laser hits or full-digitized waveforms (Riaño et al., 2004; Næsset, 2004; Andersen et al., 2005), this study extracted the LiDAR predictors, i.e., CHDs or CHQs, from the LiDAR-derived CHM. However, certain implicit relationships exist between raw laser hits and CHMs, and such relationships can be modeled so that CHDs and the height distributions of laser hits could be inferred from each other under certain assumptions (Sun and Ranson, 2000; Kotchenova et al., 2003). Therefore, it is suggested that the proposed models be extended to use the height distributions of original laser hits as predictors that may include first, last or all LiDAR returns as deemed appropriated by the analysts. By doing so, one may characterize understory structures. However, a side effect is that the use of intermediate or last laser returns should invalidate the scale-invariance of the models due to the nonlinear scalability of the height distributions derived from such laser returns. In such a situation, the scale-independence is caused by the non-uniform sampling pattern of laser hits. Future studies may investigate the utility and applicability to such functional models for biomass estimation using LiDAR height distributions of raw laser hits as the predictor.

No explicit form is assumed for the allometry equation $B(D)$ in Eq. 4.7 during the derivation of the linear model. Therefore, although it is intended to predict biomass, the model can also be applied to estimate other forest structural variables such as basal

area, stem density, crown fuel weight, canopy base height and timber volume if these structural variables use the DBH as a proxy in their allometry equations. For example, assuming $B(D) = \pi D^2 / 4$, we get a model for basal area; as an extreme case, taking $B(D) = 1$, we get a model for stem density. Besides, , we strongly suggest that future studies should examine the utility of the linear models with CHDs as predictors to estimate those forest structural variables such as Leaf Area Index (LAI) that have no explicit functional relationship with stem diameters, mainly for two major reasons: first, structural variables such as LAI are linearly scalable when changing scales, and the scale-invariant linear model we proposed can guarantee this scalable property; second, a canopy or LiDAR height distribution contains more information than any subset of its extracted statistics such as mean height, quantile-based heights and canopy density metrics. As discussed earlier, investigators may also use the discretized high-dimensional data of CHQs as predictors with a machine learning technique for developing an implicit nonlinear prediction model.

The overestimation of the overall biomass with either scanner CHM or PALS transects is attributed to several factors. First, the reference value of biomass, obtained by averaging the LiDAR-derived biomass map, only captures the biomass in those trees that were identified from CHM by the tree-finding algorithm. Due to the omission errors of the algorithm, the reference value represents an underestimate of the biomass in the vegetative components captured by CHM. This point is illustrated by the scatterplots of Figure.4.5, the lower left corners of which show a plot-level reference value of zero due to absence of identified trees, but the predicted values are small yet non-zero due to presence of some canopies in CHM. This discrepancy also leveraged all the fitted lines to the scatterplots of prediction vs. reference to tilt upward near zero biomass as shown by the solid lines compared with the 1:1 lines (Figure.4.5). Second, although we attempted to mask out the non-forest area by referring to the QuickBird image, not all the non-vegetative components are removed from CHM by doing so. Therefore, the retained man-man structures in the masked CHM contributed to the overestimation of biomass. As to the PALS profiling transects, besides the above two factors, the

overestimation may also be caused by the use of Line intersect sampling for estimation of areal attributes. Forest and non-forest patches are not homogeneously mixed over the study area; therefore, the use of systematic line sampling, especially with only few transects, could result in an overall bias (Zhao et al., 2008), which corresponds to an overestimate in our case. Overall, the PALS-based biomass estimate using all six transects is within 10% of the scanner reference value, which suggests the utility of PALS as a reconnaissance tool for quick assessment of forest resources.

4.7 Conclusions

In this chapter, we proposed the use of CHDs or CHQs as independent variables for biomass prediction, and accordingly presented two scale-invariant functional biomass models: one is a linear model that uses the whole curve of LiDAR-drived CHD as a predictor, and another is an equivalent nonlinear model that uses the CHQ as a predictor. In addition to scale-invariance, another major consideration for choosing these predictors is to incorporate as much information available in LiDAR measurements as possible into the models. Although they retain no horizontal information of LiDAR hits(x and y coordinates), CHQs or LiDAR height distributions of raw laser hits contain much more information than any subset of its extracted statistics such as mean height, truncated mean height, quantile-based height and canopy density metrics. To provide theoretical justification, we also developed an accompanied mathematical framework that helps to formulate our models based on several moderate assumptions.

The results from intensive evaluations based on the synthesized realistic datasets provide initial proofs that our models can accurately predict biomass and have consistent predictive performances across a variety of scales. Since the CHDs are obtained in the form of normalized histograms with a specified sampling height bin, we find that models with a carefully selected large height bin (e.g., less than 5.0 m) can be as effective as those with finer height bins. The results also show that in our experiments, a training sample size of around 50 or less is enough to guarantee a good fitting of the linear

functional model, as long as the training sample is representative of the forest conditions that the model will be applied to. No attempt was made to directly fit the nonlinear model, and future studies may examine the effectiveness of advanced nonlinear regression techniques such as SVM and Gaussian Processes in regressing biomass on the high-dimensional data obtained from the discretization of CHQs. For further research, it is expected that the proposed models can be effectively used to predict forest structural variables other than AGBM, such as stem density, basal area, timber volume, crown fuel weight, and even LAI, either from CHDs or LiDAR height distributions. Future research should also be carried out to use the models for forest inventory tasks where analysis units vary in size and shape. To augment the applicability of our models, it is suggested that auxiliary information from multispectral imagery should be integrated with LiDAR data for developing strata-specific models.

CHAPTER V
LIDAR-BASED MAPPING OF LEAF AREA INDEX AND ITS COMPARISON WITH MODIS LAI PRODUCTS IN AN EASTERN TEXAS FOREST

5.1 Introduction

As a key canopy structural characteristic, Leaf Area Index (LAI) serves as important input or state variable for a variety of process-based ecological and biogeochemical models, especially for those that involve modeling the exchanges of energy and mass at the atmosphere-land interface, or the photosynthesis and respiration of vegetation for carbon cycling simulation (Turner et al., 2004). LAI is typically defined as the total one-sided area of green foliage per unit ground surface (Chen & Black, 1992). Both direct and indirect techniques, e.g., destructive sampling and optical methods, exist for collecting in-situ LAI measurements (Jonckheere et al., 2004). These in-situ techniques, however, are impractical for measuring LAI over large areas due to the prohibitive costs (Cohen et al., 2003). Instead, researchers often resort to remote sensing for spatially-explicit mapping of LAI at landscape or regional levels. Reliable and accurate estimation of LAI, therefore, has become a primary task in exploiting the potential of remotely-sensed data for biophysical variable retrieval, as demonstrated both by early work in using optical imagery for LAI, and more recent efforts in estimating LAI with ranging measurements of LiDAR (LIght Detection And Ranging) (Lefsky et al., 2002).

The foundation for optical remote sensing of LAI is based on the spectral responses to LAI changes. The strong relationship between LAI and some vegetation indices is observed from experimental data, and also revealed theoretically by physical-based canopy reflectance models (Myneni et al. 1997; Eklundh et al., 2001). Among others, the utility of NDVI for estimating LAI has been proven by the extensive studies across various biomes using different remote sensing datasets such as Landsat TM/ETM+, MODIS, AVHRR. It is found that LAI-NDVI relationships not only depend

on vegetation types but also vary seasonally and annually (Wang et al., 2005). Many studies report that NDVI saturates with high LAI, particularly over deciduous forests (Birky,2001). At present, MODIS LAI layers at a resolution of 1 km are operationally produced by means of a radiative transfer-based algorithm together with an LAI-NDVI backup scheme. The accuracy of these products is of major concern to the scientific community. In a study over broadleaf forests, Shabanov et al. (2005) concluded that the precision of MODIS reflectance, the natural variability of surface spectrum and the mixture of species usually set a limit on the accuracy improvement of LAI retrieval. In practice, to guide the informed use of MODIS LAI products, a lot of validation effort has been done or is ongoing at a range of sites worldwide. Previous results suggested that there often exists an overestimate in the MODIS LAI of Collection 3 as compared to in-situ values or the extended estimates from relatively high-resolution imageries (Wang et al., 2004; Cohen et al., 2006). The major recognized difficulties in assessing the quality of MODIS LAI include the limited number of representative validation sites and the scale discrepancies between in-situ and MODIS measurements, although other factors, such as the uncertainties in radiometric correction of MODIS surface reflectance and the mis-registration of MODIS data with reference data, also will complicate the validation processes.

Recent advances in airborne laser scanners (commonly known as LiDAR) bring a breakthrough in canopy remote sensing, with an enhanced capability of direct characterization of canopy vertical structures. A relatively high probability of laser penetration into canopy allows for better characterization of understories, and at the same time alleviates the saturation problem of optical remote sensing for large LAIs or biomass. The body of LiDAR literature on ecological and environmental studies is growing in such aspects as mapping terrain topography and estimating biophysical parameters (e.g., biomass, canopy density, LAI and fuel parameters) at various analysis units (e.g., individual tree, plot, stand, and woodland) (Brandtberg et al., 2003; Lim et al., 2003; Holmgren, 2004; Popescu and Zhao, 2008). A key factor concerning the above-tree level estimation with LiDAR is the choice of appropriate LiDAR metrics

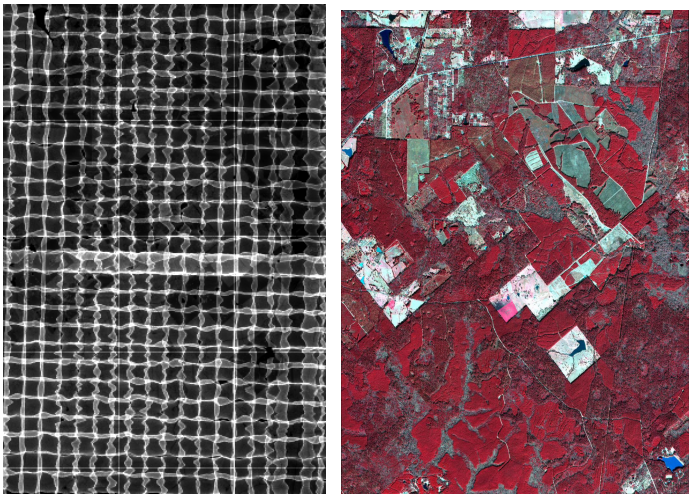
(predictors) in an effective model form, preferably with certain physical meaning. LiDAR metrics that have been previously investigated for LAI mainly include mean height, maximum height, percentile height, height of median energy, and percentage of certain type of LiDAR hits (ratio metrics). In particular, the ratio metrics such as Laser Penetration Index (LPI) and Laser Interception Index (LII) prove effective in estimating LAI (Barilotti et al., 2006). Kusakabe et al. (2005) also suggested the use of "mean free path" (penetration length into canopy) as a proxy for estimating LAI. Like optical remote sensing of LAI (Tian et al., 2002; Sarriguesa et al., 2006), estimating LAI by LiDAR is also subject to scale issues, which include but are not limited to the choice of an "optimal" resolution at which to build regression models as well as the scaling-up/down problems due to the scale-independence of the regressed models (Patenaude et al., 2004). Though not indicated explicitly, almost all the LiDAR LAI models investigated previously are scale-dependent, which is exemplified by the fact that the estimated LAI of a region by using a single metric extracted for the region does not equal the aggregated value of estimations over the sub-regions that partitions the region (Patenaude et al., 2004).

Our primary goal of this study is to continue exploring the capability of discrete-return LiDAR for spatially-explicit mapping of LAI, with a secondary goal to examine the consistence of our LiDAR-derived LAI map with MODIS standard LAI products over an eastern Texas forest. We addressed the following specific sub-problems: (1) to investigate the effectiveness of a set of LiDAR metrics, including several newly proposed ones, (2) to propose the use of laser height distribution (LHD) as predictor in a linear functional LAI model, (3) to assess the utility of integrating multispectral imagery (i.e., QuickBird) with LiDAR for improving the LAI estimate accuracy, and (4) to scale up the LiDAR LAI map for comparison with MODIS products.

5.2 Materials

5.2.1 Study Area

The study area is a 48-km^2 forested region in the eastern Texas of the southern U. S. (30° 42' N, 95° 23' W). The area mainly comprises pine plantations in various developmental stages, old growth pine stands in the Sam Houston National Forest with many of them having a natural pine stand structure, and upland and bottomland hardwoods. The major species include Loblolly pines (*Pinus taeda* L.) and deciduous trees such as water oak (*Quercus nigra* L.), red oak (*Quercus falcata* Michx), sweetgum (*Liquidambar styraciflua* L.), and post oak (*Quercus stellata* Wangenh.). Much of the southern U.S. is covered by forest types similar to those of our study area, with similar species, productivity and patterns of land use and land cover. The area is characteristic of a gentle topography; its elevation varies from 62 m to 105 m with an average of 85 m.



(a) (b)

Figure 5.1. The study area, i.e., a 48-km^2 forested region in the eastern Texas of the southern USA: (a) The pattern of scanning LiDAR swaths, and (b) a Quickbird mulitspectral image over the study area.

5.2.2 Hemispherical Photographs and In-situ LAI

Field work was undertaken from May to July in 2004. Hemispherical photographs were taken at 53 circular plots established across the study area, with plot centers geo-referenced by a differential GPS. The photographs have a resolution of 3264 × 2448 pixels, and were captured at 1.5 m above ground using a horizontally-leveled CoolPix 8700 digital camera (Nikon) equipped with an FC-E9 fisheye lens converter (Nikon). Other ground inventory data such as tree height, crown width, and crown class, were also tallied but will not be used in the present study.

5.2.3 LiDAR Dataset

The ALS data were acquired during the leaf-off season in March 2004 with an Leica-Geosystems ALS40 flying at an average elevation of 1000 m, by M7 Visual Intelligence of Houston, Texas. The LiDAR system was operated to record two returns per pulse, i.e., first and last, with a reported horizontal and vertical accuracy of 20-30 cm and 15 cm, respectively, and was configured to scan +/- 10 degree from nadir, resulting in a swath of about 350 m wide on the ground. The dataset features a full coverage from either of two perpendicular directions, with 19 flight lines in the north-south direction and 28 in the east-west direction, resulting in an average of 2.6 laser hits per m^2. From the raw LiDAR points, a Digital Elevation Model (DEM) was derived using a proprietary ground-filtering package by the data vendor; a Canopy Height Model (CHM) at a resolution of 0.5 m was created by first interpolating the canopy hits that are the laser hits of maximum height on 0.5 x 0.5 m cells and then subtracting the DEM from the interpolated canopy surface. Furthermore, in reference to the DEM, the z coordinate of the original LiDAR hits were registered locally relative to the ground so that ground hits have a z of zero. All the LiDAR-related data was georeferenced in a UTM 15N coordinate system with the WGS84 datum.

5.2.4 QuickBird Multispectral Imagery

A Quickbird (Digital Globe, Inc.) scene acquired in 2004 is available over our study area. The image has a spatial resolution of 2.4 m with four spectral bands, i.e., blue (450-520 nm), green (520-600 nm), red (630-690 nm), and NIR (760-900 nm). Radiometric calibration and ortho-rectification were applied to the image by our data vendor. The coordinate system used is the UTM 15N with WGS84 datum. An examination of 10 conspicuous feature points revealed that the image and LiDAR CHM geographically registers well with an error of less than 2.4 m.

The purpose of incorporating the Quickbird image into this study is to extract thematic information for distinguishing forest types as well as to assess the utility of NDVI for estimating LAI when integrated with LiDAR. For the first purpose, we applied the maximum likelihood classifier to the image for mainly differentiating pines, hardwood, mixed forests, and grassland. The classification produced an overall accuracy of 86.5% through an on-screen evaluation of a random subset of 200 testing pixels.

5.2.5 MODIS LAI Products

A series of Terra MODIS 8-day composite LAI products from the Julian day of 177 to 241 in 2004 were obtained from the Earth Observing System data gateway. These are Collection 4 products that are retrieved from MODIS surface reflectance products by a radiative transfer-based LUT algorithm. Due to various factors that influence the retrieval process, all the LAI products contain a quality flag layer indicating the LAI quality and cloud states. Such status information, for example, includes which algorithm, the main or backup, is invoked, whether the retrieval is conducted under “saturation condition”, and so on. The products are delivered at a resolution of 1 km in the Integerized Sinusoidal Projection (ISP). Our co-registration schemes for overlaying MODIS products with LiDAR data will be detailed in a later section.

5.3 Methods

5.3.1 In-situ LAI from Hemispherical Analysis

To better understand the uncertainties in ground LAI estimates, we considered two factors in hemispherical analysis, i.e., analysts and LAI algorithms. As such, we first employed two experienced analysts to perform binary segmentation on the hemiphotos into sky or obscured pixels using HemiView (Delta-T Devices Ltd., UK, 1999), for the purpose of evaluating the subjectivity in hemispherical analysis due to the manual and interactive selection of segmentation threshold values. Then, we applied two LAI methods to the segmented hemiphotos in order to assess the sensitivity of LAI estimates to algorithms.

The two methods, though with slightly different assumptions, are all based on "gap-fraction" formula of the Beer's law in a canopy,

$$\tau(\theta,\varphi) = \exp[-G(\theta,\varphi)L/\cos(\theta)] \tag{5.1}$$

where L is LAI; $\tau(\cdot)$ represents the gap fraction in the direction of (θ,φ) where θ is the zenith angle and φ the azimuth angle; and G is the Ross-Nilson G-function that refers to the fraction of a unit foliage area projected onto the plane normal to (θ,φ). Eq. 5.1 assumes that foliage elements are distributed randomly and independently. From the segmented hemiphotos, we obtained estimates of $\tau(\theta,\varphi)$ over some annuli or sectors, which will be used to calculate LAI according to Eq. 5.1 with the following two methods.

Method 1

This method makes use of a specific leaf angle distribution, i.e., the ellipsoidal distribution, which gives,

$$G(\theta;x) = \frac{\sqrt{x^2\cos^2\theta + \sin^2\theta}}{x + 1.774(x+1.182)^{-0.733}} \tag{5.2}$$

where the leaf distribution is supposed to be azimuthally symmetrical so that G only depends on zenith angles, and x is a constant parameterizing G. Assuming Eq. 5.2, LAI

and x can be simultaneously estimated by an iterative optimization procedure in terms of minimizing the squared error between the gap fractions t observed from hemiphotos and those predicted theoretically from the combination of Eqs. 5.1 and 5.2, as implemented in HemiView (Delta-T Devices Ltd., UK, 1999). In this study, the observed gap fractions τ for this method were estimated over 18 x 8 sky sectors with 5 and 45 degrees for zenith and azimuth divisions, respectively.

Method 2

In Method 2, we also assume that G is azimuthally-independent, but do not restrict its functional form. The method, instead, is based on the Miller's theorem

$$\int_0^{\pi/2} G(\theta)\sin\theta \mathrm{d}\theta = 0.5 \tag{5.3}$$

which, combined with Eq. 5.1, gives an LAI formula,

$$\begin{aligned} L &= -2\int_0^{\pi/2} \ln[\tau(\theta)]\cos\theta\sin\theta \mathrm{d}\theta \\ &= \int_0^{\pi/2} l_w(\theta)\mathrm{d}\theta \end{aligned} \tag{5.4}$$

where for notational convenience, the integrand has been denoted by

$$l_w(\theta) = -2\ln[\tau(\theta)]\cos\theta\sin\theta \tag{5.5}$$

However, two practical difficulties exist when applying Eq. 5.4 to the observed $\tau(\theta_i)$ for LAI L. First, some values of the observed τ, especially on annuli at large zenith angles, tend to be zero, thus causing numerical overflows in logarithm evaluations $\ln[\tau(\theta_i)]$. A remedy for this problem in previous studies is to simply add 1.0 or 0.5 pixel of openness to these problematic annuli or sectors; such addition, however, appears to be artificial (van Gardingen et al., 1999). Second, an appropriate quadrature procedure is needed to evaluate the integral of Eq. 5.4 based on the gap fractions $\tau(\theta_i)$ observed at only several zeniths θ_i's. To solve these two difficulties, in Method 2 we fit a nonparametric curve using the smoothing spline to all the observed $l_w(\theta)$'s except the problematic ones that have zeros of $\tau(\theta_i)$; by doing so, $l_w(\theta)$ can be estimated at any θ, including the θ_i

associated with problematic annuli. Then, the resulting curve is integrated to estimate LAI according to Eq. 5.4, (Figure 5.2)

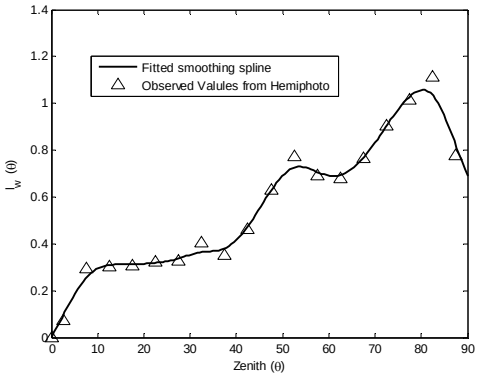


Figure 5.2. An illustration of Method 2 for calculating LAI with Eq. 5.4

The effects of analysts and methods on the in-situ estimates of LAI were analyzed with the Repeated measures ANOVA analysis. Due to the lack of ground-truth LAI data of higher accuracy, we fail to determine which of the four sets of LAI (two analysts by two methods) is more accurate. As a compromise, we instead used the average of the four sets in the subsequent analysis for relating the in-situ LAI with LiDAR metrics in the model development.

5.3.2 Parametric LAI Models Using LiDAR

Analysis of the LiDAR Data for Model Development

A better understanding of the LiDAR data characteristics helps to extract more meaningful LiDAR metrics. Our LiDAR dataset, collected by a first/last return scanner, consists of three types of laser hits: single returns (SR), which correspond to those of pulses that produce only one echo; first returns (FR) and last returns (LR) which are the first and last echoes of the pulses that have multiple echoes. In the later presentation, we strictly discriminate SRs from FRs, although SRs can be loosely deemed as either FRs or LRs. In terms of the targets intercepted, laser hits can be categorized into either ground

or canopy hits: Ground hits are often identified as those that are below a prescribed height threshold so that they include not only the true "ground" hits of zero height but also near-ground understory hits; canopy hits can be further divided into those from crown surfaces and those from inside or below crowns (Figure 5.3).

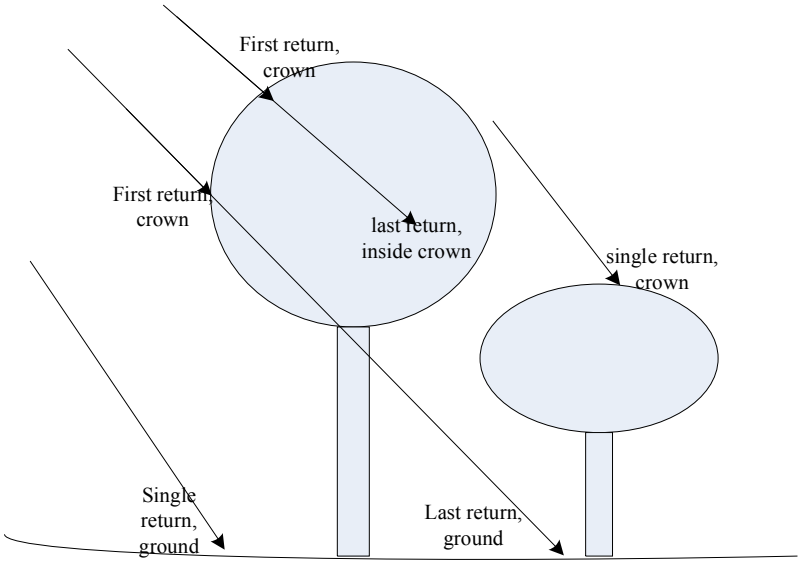


Figure 5.3. Classification of LiDAR hits into single, first and last returns that could be reflected back from different layers of canopies, e.g., crown surfaces, inside or below top crowns, and ground. Note that the arrows indicate the locations of laser hits.

For conceptual conveniences, we presume that crown-surface hits (CSH) can only be SRs or FRs; inside- or below-crown hits (IBCH) can only be LRs, and ground hits (GH) can only be SRs or LRs. The last presumption is based on the consideration that a pulse rarely produces two distinct echoes within a short distance near the ground so that no pulse can have both FR and LR coming from the ground. The above classification scheme of laser hits is also illustrated in Figure 5.3., In addition, the following relationships should roughly hold as to the numbers of each type of laser hits over an area:

$$
\begin{aligned}
N_{fr} &= N_{lr} \\
N_{pulse} &= N_{sr} + N_{lr} = N_{sr} + N_{fr} \\
N_{total} &= N_{sr} + N_{fr} + N_{lr} = N_{sr} + 2N_{fr}
\end{aligned}
\tag{5.6}
$$

where N_{pulse} and N_{total}, i.e., numbers of laser pulses and laser hits, are different in that a pulse with two returns is counted only once in N_{pulse} while it is counted twice in N_{total} due to the two returns; the other subscripts in Eq. 5.6 should be self-explanatory. Of particular note, the relationships only roughly hold, due to the boundary effects that the first return of a pulse does hit within the area but its last return is outside the boundary.

LiDAR ancillary data such as the intensity and scanning-angle provide additional information which may be of practical value for specific applications. For example, Hopkinson and Chasmer (2007) recently examined an intensity-based ratio metric, i.e., the ground power versus the total power, and found it an effective predictor for canopy gap fractions. According to a single-scattering model for "hot-spot" viewing geometry, the reflected laser intensity from a volume dV over crown surfaces is formulated as,

$$\Delta I = I_0 u_0(z) R_l G(\theta) dV \tag{5.7}$$

where I_0 is the incident laser intensity; $u_0(z)$ is the leaf area density defined over crown surfaces at a location z; R_l is the reflectance of leaves which are assumed to be Lambertian; and $G(\theta)$ is the G-function as defined in Eq . 5.1. Eq. 5.7 shows that the intensity of returns from crown surfaces is proportional to leaf area densities. Be aware, however, that I_0 varies from pulse to pulse, depending on sensor-target distances, and therefore, flight altitude and scan angle; and that LiDAR intensity measurements of most currently commercial ALS are not calibrated, which somehow limits the practical use of LiDAR intensity data for quantitative applications, e.g., estimating leaf area densities.

On the other hand, the scanning angle of a pulse, recorded by an ALS, may be different from the local look angle θ_i that is defined as the angle between the pulse incident direction and the local normal to the ground. This disparity is due to the instability of flying platforms and the rugged topography. If needed, the local look angle q_i for a pulse with two returns can be calculated from the coordinates of the first and last

returns by,

$$\cos\theta_i = \frac{|\, z_{i,lr} - z_{i,fr} \,|}{|\sqrt{(x_{i,lr} - x_{i,fr})^2 + (y_{i,lr} - y_{i,fr})^2 + (z_{i,lr} - z_{i,fr})^2}} \tag{5.8}$$

where the subscripts i, fr and lr refer to the pulse index, first return and last return, respectively; z is the height after the terrain is subtracted. For a single-returned pulse, θ_i can be approximated by considering its nearest two-returned pulse.

Based on the above analysis, we examined a variety of LiDAR metrics to be related with LAI. These metrics include penetration-based indices and canopy height metrics, some of which have been examined previously. Noteworthily, most of the metrics are chosen and constructed simply on the basis of heuristic clues rather than analytical and physical evidences. Therefore, although our choices possess certain physical meaning, their prediction abilities need to be justified only by experiential performances when regressed against in-situ LAI.

Laser Penetration Metrics (LPM)

The tenet of using penetration metrics is that, by analogy to gap fractions of the Beer's Law in Eq. 1, the percentage of penetrating laser hits provides an indication of the density and amount of foliage. However, the proper form of a laser-based ratio percentage is unclear because the LiDAR measurements, observed with a finite-size beam laser spot, reflect the collective responses of an ensemble of leaves while field measurements with a thin beam, e.g., MacArthur and Horn (1969), represents observations at smaller scales.

To construct penetration ratios, we used three types of penetrating hits as numerators: ground hits (grd), ground plus inside-canopy hits (in+grd), and ground hits only of single returns ("sr^grd"); also, we used two types of total number of laser hits as denominator: N_{total} that corresponds to all hits where a pulse with two returns are

counted twice, and N_{pulse} that corresponds to incoming pulses or first returns. Consequently, a total of six ratios were produced that are presented respectively by,

$$
\begin{aligned}
r_{grd/total} &= \frac{N_{grd}}{N_{total}} = \frac{N_{sr\^{}grd} + N_{lr\^{}grd}}{N_{sr} + 2N_{fr}} \\
r_{grd/pulse} &= \frac{N_{grd}}{N_{pulse}} = \frac{N_{sr\^{}grd} + N_{lr\^{}grd}}{N_{sr} + N_{fr}} \\
r_{in+grd/total} &= \frac{N_{in} + N_{grd}}{N_{total}} = 1 - \frac{N_{sr\^{}cs}}{N_{sr} + N_{fr}} \\
r_{in+grd/pulse} &= \frac{N_{in} + N_{grd}}{N_{pulse}} = 1 - \frac{N_{sr\^{}cs}}{N_{sr} + N_{fr}} \\
r_{sr\^{}grd/total} &= \frac{N_{sr\^{}grd}}{N_{total}} \\
r_{sr\^{}grd/pulse} &= \frac{N_{sr\^{}grd}}{N_{pulses}}
\end{aligned}
\tag{5.9}
$$

As variants to the above metrics, intensity-based ratios can also be derived using the sums of intensity values rather than the numbers of hits. However, to constrain the value of ratios within [0,1], we considered only the intensity ratios that use the sum of intensity of all hits as denominators. The three resulting intensity-based counterparts are symbolized respectively as $i_{grd/total}$, $i_{in+grd/total}$, and $i_{sr\^{}grd/total}$

LPMs Adjusted by Look-angles

Penetration rates through a canopy also depend on incident directions, often with a larger incident angle resulting in less penetration as revealed in Eq. 5.1. In the LiDAR data, incident directions of laser pulses slightly oscillate from pulse to pulse between a maximum of scann angle, due to the scanning mechanism of ALS. Moreover, the displacements of different field plots relative to the flightlines are different, thus contributing to the variation in incident angles among different plots. In this study, most of our field plots were observed from multiple flight lines due to the cross-hatch flight

patterns, thus increasing the disparity of incident angles among different plots. Therefore, to compensate such variations, we take into account looking angles when designing a penetration ratio metric for LAI, and the LPMs adjusted by looking angle θ_i are given, for example, as follows,

$$r_{adjust,grd/total} = \frac{\sum_{i=0}^{N_{grd}} 1/\cos\theta_i}{N_{total}} \tag{5.10}$$

where $\cos\theta_i$ is obtained with Eq. 5.8. We adjusted for the look angle effects only for those LPMs based on pulse counts, i.e., Eq. 5.9, not for the intensity-based LPMs. The adjusted counterparts to the remaining five ratios of Eq. 5.9 can be obtained by analogy to Eq. 5.10. It is noted that the form of the formula for the adjusted LPM is not strictly physical-based.

Height-related Metrics (HRM)

LAI should be proportional to canopy volume, foliage density, or their product. For canopy volume, a LiDAR-based surrogate is LiDAR-based height metrics, e.g., mean canopy height as used in Chen et al. (2007). For foliage density, a surrogate is the "mean free path", i.e., the mean penetration length of lasers into top canopy surfaces, which is deemed to be inversely related to foliage density as argued in Kusakabe et al. (2005); another surrogate to foliage density, i.e., the average intensity of first canopy returns, could be obtained by referring to Eq. 5.7 where $u_0(z)$ is the foliage density near crown tops.

Based on the above heuristics for predicting LAI, we investigated several height metrics, including the mean height of first returns, $\overline{h}_{fr}$, the mean height of all returns, $\overline{h}_{total}$, the 50% and 80% percentile heights of all returns, $h_{50\%}$ and $h_{80\%}$, the maximum height, $h_{\max}$, and the mean CHM height, $\overline{h}_{chm}$. We also investigated both the surrogates to foliage density. Specifically, the "mean free path" is calculated based on only canopy hits with the free path of a single-return pulse being zero and that of a double-return

pulse being the distance between its first and last returns; the average intensity of canopy hits is based on only canopy hits of first return.

Statistical Analysis Relevant to LPMs and HRMs

Several apparent factors make it difficult to build models using the aforementioned LiDAR metrics to predict LAI. First, in deriving the penetration indices, an appropriate height threshold for separating ground and canopy hits needs to be determined. In several earlier studies, the threshold is chosen as the height at which the camera was set up, but such a choice guarantees no optimality. Second, exact forms of relationships between LAI with each metric are not theoretically available, although both linear and logarithmic models have been used in previous studies (Morsdorf et al., 2006). Third, it remains undetermined as to which metric or subset of the metrics provides the best prediction power. Fourth, the selection of predictors and models is affected by plot sizes at which LiDAR metrics are extracted as reported in Riaño et al. (2004) because there is not a definite plot size corresponding to the hemiphotos that were used to derive in-situ LAI. Last, forest types, i.e., pines or hardwood, may well influence the relationships between LAI and LiDAR metrics. Due to the compounding effects of the above factors as well as the limited number of ground LAI observations, it seems impractical to enumerate all the combinations of metrics, models, plot sizes, and forest types for model selection. As an expedient, the following strategy is employed to seek an "optimal" model.

When using the LPMs or the adjusted LPMs as predictors, the logarithmic model (log-model) with no intercept term is used because of its resemblance to the Beer's law; however, to accommodate possible difference between pines and hardwood, the species-stratified equation was used, i.e.,

$$L = \beta_1 \cdot \ln(x) + \beta_2 \cdot t \cdot \ln(x) + \varepsilon \tag{5.11}$$

where x stands for any one of the 15 LPMs, including six hit-number-based ones, three intensity-based ones, and six adjusted ones; t is a binary variable distinguishing pine plots from those of hardwood/mixed; β_i's are coefficients to be determined, and ε is

the error term. Due to the possible combined effects of LPM and plot size on the model performance, the model of Eq. 5.11 was examined for each LPM predictor at a series of plot sizes ranging from 5.0 m to 39.0 m in radius with a step of 1.0 m in order to choose a most appropriate scale for building the LAI model. Among the 525 fitted models, i.e., 15 LPMs by 35 plot sizes, the one with the least root mean square (RMSE) was selected, which indicates not only the best predictor but also the most appropriate plot size for the predictor. Of particular note is that during the model selection, we fixed the height threshold to be 2.56 m that was used to separate ground and canopy hits in constructing all the LPMs. This threshold, not equal to the camera setup height (1.5m), was chosen for reasons that will become clear later.

When using HRMs as predictors, two cases were investigated, i.e., single-variable and multiple-variable regression models. In the case of models using single variable, a total of 18 predictors were examined, including six height metrics and 12 product metrics (i.e., six height metrics by two foliage density proxies). The model form was chosen to be,

$$y = L = \beta_0 + \beta_1 t + \beta_2 x + \beta_3 x \cdot t + \varepsilon \tag{5.12}$$

where x is any of 18 predictors; t again is the binary variable used to differentiate pine plots from those of hardwood/mixed; β_i's are the regression coefficients, and ε is the error term. Furthermore, to accommodate any possible nonlinear relationships, the power transformed LAI and predictors are also examined, i.e,

$$y = L^{\lambda_1} = \beta_0 + \beta_1 t + \beta_2 x^{\lambda_2} + \beta_3 x^{\lambda_2} t + \varepsilon \tag{5.13}$$

where λ_1 and λ_2 are exponents that do power transformation to LAI and the predictor, respectively, and all the other terms are the same as in Eq. 5.12. The two exponents are estimated by the bi-variate Box-cox transform methods. A zero exponent in the Box-cox transform reduces to the logarithm transform. In addition, the use of the Box-Cox method renders the regression analysis more theoretically sound. In the case of multiple regression models, the pooled set of the above 18 predictors, without power transformation, was employed as candidates to be selected into an optimal multiple

linear regression model by using the stepwise regression procedure; in this case, the model was not stratified by species (forest types), due to the relatively small sample size, especially those of hardwood plots. Of note, the effect of plot size was fixed in both cases of single- and multiple variable models, and we used the optimal plot size of the LPM-based model for extracting the HRMs. It is also noted that in this study, we made no attempt to build a model that uses both LPM and HRMs as predictors.

5.3.3 Functional Models Using LiDAR Height Distributions

LiDAR height distributions (LHD) represent the relative frequencies of laser hits at given heights, which can be obtained for an analysis unit (e.g., plots) by counting laser hits that fall into a short height-bin around a given height, and then normalizing the count with the respect to the product of the total number of hits and the length of height-bin. An unarguable advantage of LHD is that it retains as much information as possible in that most traditional metrics such as mean height, quadratic mean height, truncated mean height, quantile height, LPMs, and percentages of canopy hits all can be directly derived from LHD. As an example, the LPM $r_{grd/total}$ in Eq. 5.9 can be calculated from the LHD of all hits $p(h)$,

$$r_{grd/total} = \int_0^{h_{thr}} p(h)dh \tag{5.14}$$

where h_{thr} is a height threshold below which laser hits are classified as ground hits.

A prediction model that uses $p(h)$ as predictor should be considered as a functional model because $p(h)$, a curve with respect to h, is a function by itself. However, in practice, the values of $p(h)$ are often available at a discrete set of heights, i.e., $\{p(h_i)\}_{i=1}^{n}$, $h_i = (i-1)\cdot\Delta h$, where Δh is a height bin used to sample $p(h)$, and n is the number of height-bins used. In this study, we investigated two such functional models for LAI, both of which are in the form of,

$$L = -k\ln\left[\int_0^{h_{\max}} K(h)\,p(h)dh\right] + \varepsilon \tag{5.15}$$

where $K(h)$ is a kernel function introduced into the model; k is the extinction coefficient and it holds from Eq. 5.1 that $k = G(\theta)/\cos\theta$ where θ is the observation direction; $h_{\max} = n \cdot \Delta h$ denotes a height above which no laser hits are present; ε is the normal error term. In essence, Eq. 5.15 is the Beer's law so that $\int_0^{h_{\max}} K(h)\,p(h)dh$ represents the gap fraction observed at the bottom of a canopy. The model parameters or unknowns that need to be inferred from training data are k and $K(h)$ which are a scalar and a function, respectively.

In practice, the unknown $K(h)$ in Eq. 5.15 can take either parametric or nonparametric form. As such, the following will deal with two models, the first of which has a parameterized $K(h)$, and the second of which has a nonparametric $K(h)$. The consideration in building the models is also presented to justify the choice of $K(h)$.

Functional LAI Model with a Parametric $K(h)$

In the first model, we employed a specific parametric model of $K(h)$, namely, the logistic function,

$$K_{\log}(h; s, h_0) = \frac{1}{1 + \exp[s \cdot (h - h_0)]} \tag{5.16}$$

where s, and h_0 are the specifying parameters. As shall be seen later from the results, $K_{\log}$ is a restricted logistic curve whose maximum and minimum asymptotes are set to 1 and 0 respectively. The specifying parameter h_0 is the height associated with the inflection point, and s is the slope parameter that determines the steepness of the transition zone around the inflection point. It is interesting to notice that when s is relatively large, e.g., > 2.0, so that the logistic curve can be approximated by a step function, the model reduces to a linear one with $r_{grd/total}$ as predictor,

$$L = -k \ln(\int K_{\log}(h; \infty, h_0) \cdot p(h)dh) + \varepsilon = -k \ln(r_{grd/total}) + \varepsilon \tag{5.17}$$

where $r_{grd/total}$ is the LPM derived by using h_0 as a threshold to separate ground/canopy hits. In such a sense, the inference of the functional model with $K_{\log}(h; s, h_0)$ provides us an optimal threshold of height $\hat{h}_0$ for determining ground hits.

The parameters to be inferred for this model includes the extinction coefficient k, and the $K_{\log}$'s specifying parameters s and h_0. These three parameters can be estimated with gradient-based optimization methods in terms of minimizing a weighted mean square error; and a conjugate gradient method was employed for this purpose. Details on the gradients used by this method are relegated to a later publication.

Functional LAI Model with a Nonparametric $K(h)$

Instead of parametrizing $K(h)$, a nonparametric curve for $K(h)$ can be used in the functional LAI model. As a result, the model inference must yield an estimate for the whole curve of $K(h)$. Moreover, it is usually beneficial to apply some constraints on $K(h)$ according to some physical heuristics. Next, we describe the choice of constraints and provide some justification for the choice as well.

Previous studies reported the use of LiDAR height distribution $p(h)$ for estimating several forest stand characteristics. Intuitively, $p(h)$ is similar to a waveform that is measured by large-footprint LiDARs because both of them represent the vertical structure of canopies, but in fact, they are obtained in different manners and are not equivalent to each other. Therefore, $p(h)$ is called a pseudo-waveform by some researchers. On the other hand, both $p(h)$ and a large-footprint waveform are different from the gap probability of the canopy $P(h)$, and they can be treated as some observed responses of LiDAR to the gap probability. Therefore, to retrieve $P(h)$, the LiDAR waveform or $p(h)$ need to be "de-convoluted" back. For example, a commonly used technique to correct waveform for gap probability $P(h)$ is the Macarthur-Horn method (Lefsky et al., 2003). For convenience of model development, we assume that the gap probability of a canopy $P(h)$ can be linked with the pseudo-waveform $p(h)$ of scanner data can with the help of a de-convolution kernel $C(x, y)$,

$$P(h) = \int C(x,h)\, p(x) dx \tag{5.18}$$

Hence, the gap probability at $h = 0$ (the bottom of the canopy) can be written as,

$$P(0) = \int C(x,0)\, p(x) dx = \int K_0(h)\, p(h) dh \tag{5.19}$$

where for notational convenience we denote $K_0(h) = C(h,0)$. By the Beer's law, i.e., $L = -k \ln[P(0)]$, LAI can be calculated as,

$$L = -k \ln[\int K_0(h)\, p(h) dh] \tag{5.20}$$

which exactly is the functional model given in Eq. 5.15, and $K_0(h)$ is the nonparametric curve that needs to be estimated.

To ensure that $P(0)$ of Eq. 5.19 falls into the interval of [0, 1] for any possible $p(h)$, $K_0(h)$ cannot be arbitrary and must meet some conditions. For simplicity, we only present some sufficient conditions for this, but they may be not necessary. First, we put a constraint that,

$$0 \le K_0(h) \le 1 \tag{5.21}$$

Second, we restrict $K_0(h)$ to be a non-increasing function. This constraint is based on the facts that $p(h)$ and $P(h)$ bears some resemblance to each other, and $P(0)$ is essentially a weighted sum of $p(h)$; therefore, it is sensible to assume that the contribution of $p(h)$ to $P(0)$ at a large height should be less than that at a small height, i.e.,

$$K_0(h_1) > K_0(h_2) \text{ if } h_1 < h_2 . \tag{5.22}$$

Unlike the parametric function model, the unknowns in this nonparametric models consist of the extinction coefficient k and the whole curve of $K_0(h)$. In addition, the model inference is subject to the constraints applied on $K_0(h)$, as listed in Eqs. 5.21, and 5.22. In this study, we used a hybrid iterative constrained and weighted linear regression technique to infer k and $K_0(h)$.

5.3.4 LiDAR-derived LAI Map

An appropriate plot size was identified by examining the overall model performances at various plot sizes, and it then was used to determine the spatial resolution at which to generate the LiDAR-based LAI maps. Specifically, the pixel size of LAI maps should be approximately equal to the plot area. Even at this identified plot size, there are a variety of models evaluated, so only those with good performances were selected to be applied for producing the LiDAR maps.

5.3.5 Registration between LiDAR and MODIS LAI Maps for Comparison

Difficulties exist when attempting to compare a LiDAR-based LAI map with that of MODIS. First of all, co-registration between LiDAR and MODIS maps must be first performed in order to compare them; however, the two types of maps, both being a raster one, are not only generated at different resolutions with the MODIS LAI maps having a 1 km^2 pixel size that is much larger than that of the LIDAR maps, but also, are geo-referenced in different coordinate systems. The registration of a MODIS LAI map in reference to a LIDAR one may be problematic because this process needs to interpolate the original MODIS pixels to obtain new values at a re-organized grid of 1km^2 pixels, and positional errors may also be exaggerated during this process due to the large MODIS pixel size relative to the LiDAR map. On the other hand, even if the re-mapping and interpolation processes contribute to no mis-registration, the original positional errors in the two sources of data can make it impractical to do pixelwise comparison (Tan et al, 2003). For this reason, no pixel-wise comparison is made in this study; instead, the statistics of LAI values over the study area, e.g., mean and variance, were referred to for a patch-level comparison. In addition, unlike most previous studies that re-projected one map to another map with the help of re-sampling, we transformed only the extent of the study area, i.e., a rectangle in UTM 15N, from the UTM 15N projection of the LiDAR map to the ISP of MODIS. The resulting extent in the ISP is a parallelogram and was overlaid with MODIS maps to determine the MODIS pixels that fall into the study area (Figure 5.4). The MODIS pixels intersecting the boundary of the

parallelogram were included for comparison only if more than 1/3 of the pixel area falls within the boundary.

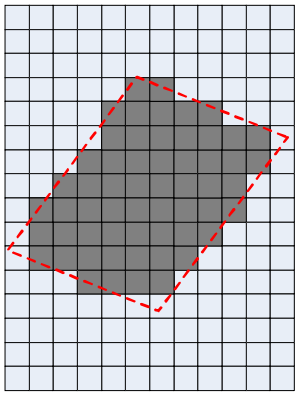


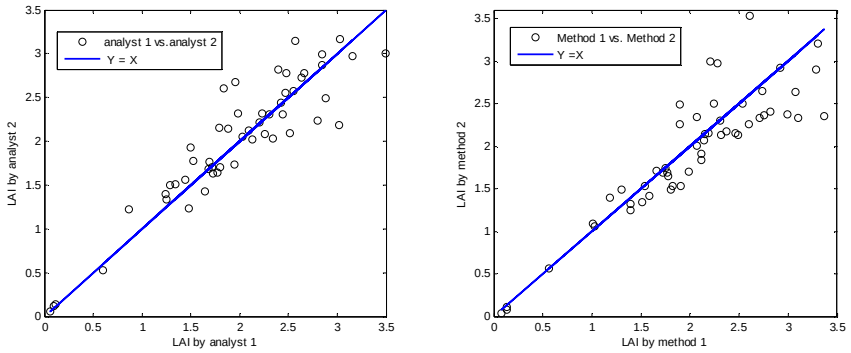
Figure 5.4. Registration scheme for determining the MODIS pixels that fall within the boundary of the study area (red parallelogram).

5.4 Results and Discussion

5.4.1 Estimates of In-situ LAI

The effects of analysts and hemiphoto analysis methods on the in-situ LAI estimates were examined. Figure 5.5 presents the comparison of LAI values obtained by two analysts as well as the comparison between the two methods. To avoid plethora of figures, the scatterplots in Figure 5.5 refer to the mean LAI values averaged over the two cases (levels) of a factor when making comparison between the two cases of another factor. It is observed that both factors affect the estimation of LAI. In terms of the strength of correlation between the two cases of each factor, the effect of methods seems more distinct than that of analysts, as also revealed by the scattering patterns of two factors. According to a Repeated measures ANOVA test that used analysts and methods as within-subject factors, and species as between-subject factor, both analysts and methods have no significant influences on estimating mean in-situ LAI values at a level of 0.05; however, the test did show that the factor of methods (p-value: 0.09) appears to

affect the LAI estimation for both species more than the factor of analysts (p-value: 0.36).



(a) (b)

Figure 5.5. Scatterplots of in-situ LAI estimates (a) between two analysts, and (b) between two hemiphoto analysis methods

For each factor, two root mean square differences (RMSD) were first obtained from the four groups of LAI estimates by taking the differences between the two levels of the factor, at each level of another factor; and the average of two RMSD (ARMSD) is used to represent the variability caused by that factor. For example, ARMSD for the factor of analysts was calculated by,

$$ARMSD_{analyst} = \frac{\| L_{a1,m1} - L_{a2,m1} \| + \| L_{a1,m2} - L_{a2,m2} \|}{2} \tag{5.23}$$

where $a1$ and $a2$ represents the two analysts; $m1$ and $m2$ represents the two methods; and $\|\cdot\|$ denotes the RMSD. The ARMSD for analysts and methods are 0.318 and 0.178 respectively; and in such a sense, it appears that analysts have more influences on the uncertainties in the in-situ LAI estimates, differently from what the Repeated measures ANOVA suggested.

5.4.2 LiDAR-based LAI Estimates Using LPMs

Four of the 15 LPM predictors were identified to consistently have good prediction abilities across different plot sizes in terms of R^2 as well as the root mean square errors (RMSE), and they are $r_{in+grd/total}$, $i_{in+grd/total}$, $R_{grd/total}$, and $r_{grd/total}$, respectively. The prediction abilities of these four LPMs are very close to each other with $r_{in+grd/total}$ slightly better than the others and $r_{grd/total}$ yielding the least R^2 among the four. All these four LPMs are far more effective than the remaining 11 predictors when using the log-model; for example, a comparison among the R^2 values of 15 LPMs for a plot radius of 20 m is depicted in Figure 5.6.

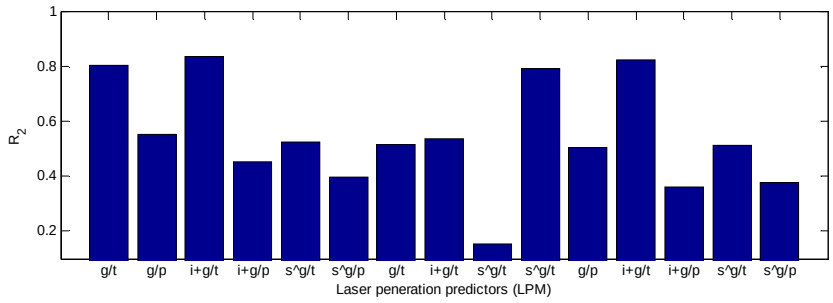


Figure 5.6. R^2 of the fitted models for different LPMs that are extracted using a plot size of 20 m. For ease of displaying, the names of LPM predictors are shortened in such a way that "g", "i", "s", "t", and "p" represent ground, inside crown, single return, total, and pulse, respectively. The first six columns are hit-number-based LPMs, the next three are intensity-based ones, and the last six are looking-angle adjusted LPMs.

The size of plots used to extract LPMs also affects the model behaviors. Our results show that for all predictors, a range of plot radii from 12 m to 30 m were generally appropriate choices for extracting LPMs to be related with in-situ LAI of hemiphotos. Over this range of plot size, only slight fluctuations in R^2 were observed, as shown in Figure 5.7 where changes in R^2 with plot size are plotted for the best four LPMs. It is noted that when using too small a plot, e.g., less than 8 m in radius, the fitted

models degraded greatly with a relatively large decrease in R^2, e.g., from 0.78 at a radius of 8 m to 0.63 at a radius of 6 m for the predictor $r_{grd/total}$.

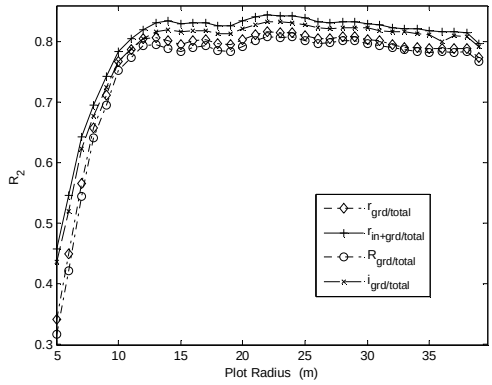


Figure 5.7. The effects of plot size on the model fitting for the best four LPMs.

5.4.3 LiDAR-based LAI Estimates Using HRMs

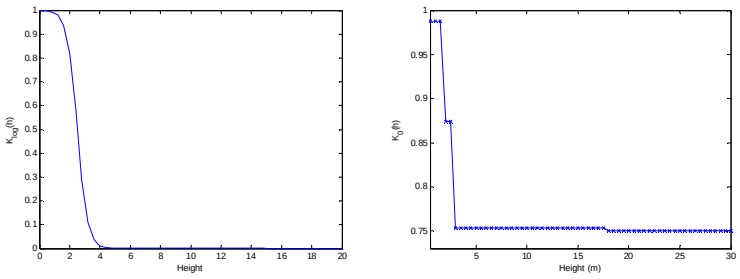
Less success was gained in using HRMs to predicted LAIs. A range of R^2 from 0.067 to 0.302 for the untransformed single-variable linear models, was observed using the training data extracted at a plot size of 20 m in radius, and the predictor with the largest R^2 0.30 is the mean height of first returns. The two product predictors resulting from the multiplication of mean height of first returns with the two foliage-density surrogates did not improve the prediction abilities. On the other hand, the transformed single-variable models using Box-cox methods have a similar range of R^2 from 0.071 to 0.302 with only marginally or no improvements in most cases. Interesting enough, for the multiple-variable models, the stepwise linear regression procedure resulted in the following model with two predictors $\frac{h_f}{l_{free}}$ and $\frac{h_a}{l_{free}}$,

$$\hat{L} = 0.4 + 2.87\frac{h_f}{l_{free}} - 2.58\frac{h_a}{l_{free}} = 0.4 + (2.87h_f - 2.58h_a)\cdot\frac{1}{l_{free}} \tag{5.24}$$

which has an adjusted R^2 of 0.75. Because $1/l_{free}$ is used as a proxy to foliage density, $(2.87h_f - h_a 2.58)$ then can be considered as a proxy to the canopy volume.

5.4.4 LiDAR-based LAI Estimates Using Functional Models

The plot size of 20 m in radius was chosen to extract LiDAR height distributions (LHD), and the extracted LHDs were used to train the two functional LAI models. The estimated parametric $K_{\log}(h)$ has an inflection point at $h = 2.56$ m as shown in Figure 5.8. This height provides an optimal threshold to separate ground and canopy hits, and in fact, it has been used in deriving all the LPMs. The estimated nonparametric $K_0(h)$ is plotted in Figure 5.8. An examination on the two estimated $K_0(h)$ curves reveals that $K_0(h)$ has more weighting on $p(h)$ at large heights than $K(h)$, due to the heavy tail of $K_0(h)$.

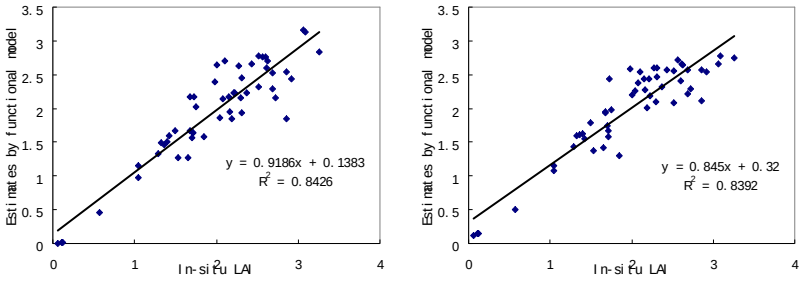


(a) (b)

Figure 5.8. The estimated curves of parametric $K_{\log}(h)$ and nonparametric $K_0(h)$ for the functional LAI models. The LiDAR height distributions used to estimate these curves are extracted at a plot size of 20 m in radius.

Figure 5.9 depicts the scatterplots of estimated LAI versus in-situ LAI, for the functional model with the nonparametric $K_0(h)$ and the best LPM model with the predictor $r_{in+grd/total}$, respectively. Although it appears that the estimates from the LPM model are better correlated with the in-situ values (R^2=0.843) than those from the

functional model (R^2=0.839), the functional model has a slightly smaller RMSE of 0.294 as compared to 0.298 of the LPM model. In addition, the estimates of extinction coefficients in the two cases differ greatly, with k being 0.1 and 0.45 for the functional and LPM models, respectively.



(a) (b)

Figure 5.9. Scatterplots of estimated LAI versus in-situ LAI, for (a) the functional model with the nonparametric $K_0(h)$ and (b) the best LPM model with the predictor $r_{in+grd/total}$.

5.4.5 LiDAR-derived LAI Map and Its Comparison with MODIS Products

Figure 5.10. LiDAR-derived LAI map with a spatial resolution of 35 m; it was generated by using the LPM predictor $r_{in+grd/total}$.

The LPM model with $r_{in+grd/total}$ as predictor was chosen to be applied to the whole study area for generating a LiDAR-derived LAI map (Figure 5.10), and the spatial resolution of the map is 35 m X 35 m which is determined by equaling the area of a pixel to that of a plot 20 m in radius. Although the QuickBird classification map was available to provide species information, the LPM model used was not stratified by species because stratification did not significantly improve the model performance. The LAI map was created by using the following non-stratified model,

$$\hat{L} = -2.24 \ln(r_{in+grd/total}) \tag{5.25}$$

And the resulting map is displayed in Figure 5.10.

Fifty-nine MODIS LAI pixels that fall within or on the boundary of the study area were selected for comparison with LiDAR LAI estimates. A series of mean and standard deviation of MODIS LAI were calculated from the 59 pixels for the time period from Julian Day of 177 to 241 with an eight day interval (Figure 5.11), and these results are compared to that obtained from the LiDAR LAI map. The mean LiDAR LAI is 1.76 with a standard deviation of 1.0; the MODIS products for the 10 dates all overestimated the mean LAI in comparison to LiDAR, and the overestimation was significant at a level of 0.05 for seven of the ten dates.

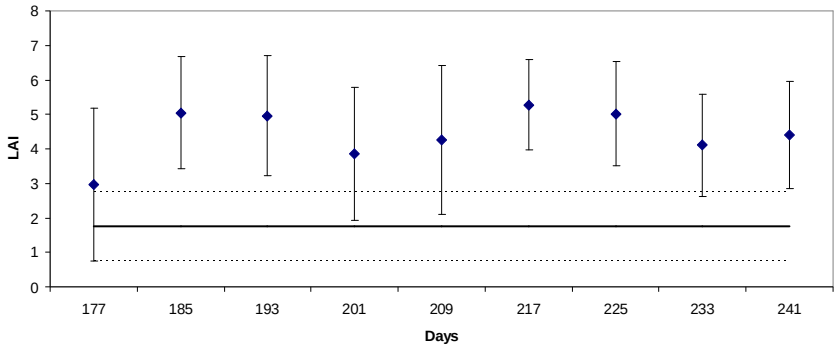


Figure 5.11. Comparison of the mean LAI over the study area between LiDAR and MODIS products on ten dates. The MODIS mean LAIs are denoted by diamonds with bars of standard deviation; the horizontal bold line is the mean LiDAR-based LAI with the two dashed lines indicating the standard deviation.

The LiDAR and MODIS LAI maps show disparate patterns in the LAI distributions. The relative frequencies of LAI for LiDAR-based and several MODIS maps are compared in Figure 5.12. It becomes clear that an upward shift exist in the MODIS LAI distributions, which indicates the overestimation by MODIS relative to LiDAR. A considerable number of pixels have zero-valued LAI in the LiDAR map while no zero LAI appears in the MODIS LAI. Only a very small portion of the MODIS pixels are below a LAI value of 3.0 whereas most of them are above 3.5; in contrast, besides the zero-valued pixels, most pixels in the LiDAR map were found within the range of 2.2 – 2.3. In addition, during the period observed, the mean MODIS LAIs over the study area vary from 3.0 to 5.3 which is 70.5% to 201.2% above that of the LiDAR LAI map.

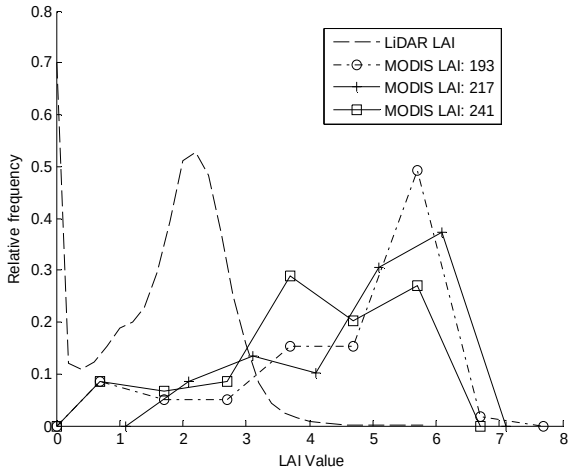


Figure 5.12. Comparison of the distributions of LAI values of the study area between the LiDAR map and the MODIS products on three dates.

5.5 Discussion and Conclusions

LiDAR measurements hold great potentials for estimating LAI accurately, in particular over medium to high biomass forests. This study demonstrates the use of several LiDAR-derived metrics as predictors for LAI over a pine-dominant eastern

Texas forest. The LiDAR predictors investigated include a variety of laser penetration indices and canopy height-related metrics as well as the height distributions of LiDAR hits. LPMs prove effective in a log-model by analogy to the Beer's law, and only partial success was achieved when using canopy height metrics. The proposed use of LiDAR height distribution as a predictor in a functional model was also shown to be feasible in estimating LAI. An unarguable advantage of a LHD is that it tends to retain as much information as possible when extracting LiDAR metrics from the raw LiDAR data, because many other commonly used predictors such as mean height and canopy density metrics can also be calculated from the LHD. On the other hand, LiDAR-derived LAI maps should be reliable enough to serve as reference data for validating the relevant LAI products generated from other sources. The LiDAR LAI map generated was compared to the MODIS LAI products at a patch level by examining the mean LAI over the study area, and it is found that the MODIS LAIs produce an overestimation as compared to that of LiDAR-derived map.

Many factors contribute to uncertainties in mapping LAI with LiDAR. First of all, the in-situ LAI data used to train a prediction model are subject to errors. The effects of analysts and methods on the in-situ estimates of LAI were examined, and the results indicate that in this study, the uncertainties caused by these factors are in the same order of magnitude as the RMSE of the fitted models. However, it seems difficult to assess the accuracy of in-situ estimates because reference LAI values with higher accuracy are not available. The discrepancies caused by the analysts are obvious because of their subjectivity in thresholding the hemiphotos, and those caused by methods result from the different assumptions in designing the algorithms. The first method of this study has a more restrictive assumption because it specifies the exact form of leaf angle distribution, which may invalidate the algorithm if the real forest conditions are far from what is assumed. A notable difficulty in developing inversion algorithm for in-situ LAI concerns the modeling of foliage clustering at different structural levels. Modeling efforts have been devoted to solving the difficulties; for example, Nilson (1999) developed formula

for inverting canopy structural variables such as LAI from gap data by explicitly considering the foliage clustering of separate crowns.

The LPM LiDAR predictors were examined only in the log-model with no intercept. Although previous studies also used some penetration indices for LAI with linear models, these models are not investigated in this study primarily due to some of their artifacts: LPM often takes values between 0 and 1; therefore, the use of LPM as a predictor in a linear model sets a limit on the dynamic range of LAI values that could be predicted, thus resulting in artificial saturation in LAI estimates; also, the intercept term, if negative, may also cause unrealistic estimates. If the intercept term is introduced in the log- models used in this study, problems may also occur; this explains why only non-intercept log-models were explored in this study. It is noted that in most cases the unrealistic LAI prediction occurs as LPM takes extreme values, and thus extrapolation may be exercised. On the other hand, some of LPM predictors such as $i_{in+grd/total}$, performed poorly when used in the no-intercepted log-model, but if an intercept term is used, the model fitting can be greatly improved; for example, when using a plot size of 15 m in radius, the R^2 increased from 0.10 to 0.65 by adding the intercept term. However, this may not necessarily justify the use of a log-model with intercept when using LPMs.

The height-related metrics, which were found useful to predict a variety of forest structural characteristics such as timber volume, crown base height and basal area, are found to be not well correlated to LAI. The products of height metrics with foliage-density LiDAR surrogates also do not improve their prediction power for LAI. In this study, HRMs were much less effective than LPMs for predicting LAI. But a multiple regression model resulting from the stepwise analysis seems useful in estimating LAI, and the variables selected into this model partially conform to the original heuristics that LAI is proportional to the product of canopy volume and foliage density. In Eq. 5.24, $1/l_{free}$ is the inverse of "mean free path", which is assumed to be positively related to the foliage density as argued in Kusakabe et al. (2005), and as such, $(2.87h_f - 2.58h_a)$ can be effectively deemed as a canopy volume metric.

The choice of an appropriate scale, e.g., plot size, for extracting LIDAR metrics is another important factor in developing LiDAR-based models. The reason for choosing the plot size is mainly because that no direct information is available about how far a camera can "see" when taking hemiphotos. Trial-and-error may be a basic method to determine an "optimal" plot size. Previous studies also provide experiential evidences on how to choose a reasonable plot size (Riaño et al., 2004). In this study, it is found that a plot radius from 10 m to 50 m were generally appropriate, and even a radius up to 100m can be used without significant degradation in the fitted model. This large upper radius limit as compared to that reported in previous studies is due to the fact that the study area, which has a large portion of pine plantations, are relatively homogeneous. Of particular note is that according to the Beer's law, the camera will see through a shorter distance under thick canopy conditions that usually have higher LAIs. The opposite is also true; for example, a hemiphoto taken under a relatively open ground with only sparse low vegetation may observe a few hundreds of meters. Therefore, a fixed plot size as used in this study may not be sufficient; future studies may investigate the use of a varying plot size for extracting LiDAR metrics during the development of models. One possible way is to examine the observed LAI values because they provide hints on how far a camera can see.

The results for the comparison of LiDAR LAI with MODIS products are consistent with previous findings that the MODIS algorithm usually overestimates LAI and the variation in MODIS LAI is often higher than that in the true values. All these findings urge the needs for possible improvement on the MODIS algorithm. At the same time, efforts are entailed to develop more appropriate procedures to validate the MODIS products, especially concerning the creation of a reference map. A commonly used approach to creating high-resolution LAI map is to rely on multispectral images to scale the in-situ LAI for a LAI map over local regions by either referring to statistical-based relationships or physical-based algorithms; one apparent drawback of this approach is the use of multispectral images which may by themselves cause problems over medium to high biomass forests. To this end, the approach developed in this study to etimate LAI

with LiDARs provides a superior alternative for generating local LAI maps as a reliable source for validating MODIS products.

CHAPTER VI
SOFTWARE IMPLEMENTATION

6.1 Introduction

Implementation of a remote sensing real-time processing system necessitates an integrated software package to coordinate hardware components for data acquisition, and more importantly to process and analyze the acquired data in-flight for information extraction. However, due to the wide variety of remote sensing systems that are designed and implemented using varied architectures and techniques, depending on the availability of hardware and the purposes of applications, no accepted protocol exists for developing software packages for real-time systems. Lienert et al. (1999), for example, developed a software system for real-time analysis and visualization of scanning LiDAR data for deriving information on marine aerosol. The software, written in Visual C++, has implemented an iterative forward stepping technique for inversion of LiDAR backscattering data. They also concluded that the software can be easily transplanted for other LiDAR systems such as Differential absorption LiDAR and oceanic LiDAR. Often enough, in some real-time systems, software is developed to be integrated with specific hardware equipment. As an example, Wu et al. (2004) developed the one-path photogrammetric program for a real-time photogrammetric mapping system in order to automatically generate Digital Elevation Models, ortho-images, and contour lines; and their computer program was developed in coordination with a digital frame camera, a position and orientation system, and an aerial survey control tool navigation system. In some cases especially when the data volume is huge and computational demands are high, core algorithms of a real-time system can be implemented directly on chips instead of being programmed by computer languages (Masayuki et al., 2002).

The choice of a language in programming for a real-time system often needs to take into consideration such factors as the compatibility between software and hardware,

the capability or functionality of a language in terms of the task needs, the cost or availability of a language to developers, and the existing version of accessible library of subroutines or procedures. Other than these major factors, the rest of the choice is up to developers' personal preferences. In this research, a scientific programming language called Interactive Data Language (IDL) is chosen to develop computer software for the proposed on-the-fly profiling LiDAR system. IDL is a so-called fourth-generation language allowing users to process data interactively. It is rich in a wide spectrum of numerical and image processing routines that are callable and apparent to users. IDL finds broad applications in the signal and image processing community. For example, ENVI, a leading remote sensing image analysis package, is developed in IDL, and being constantly extended by new IDL add-ons. Many researchers in LiDAR forestry remote sensing also develop computer procedures using IDL to handle and process LiDAR data. In this study, other reasons why IDL is preferred include the facts that IDL is platform-independent, that it allows developing event-driven graphical user interface (GUI), and that stand-alone IDL packages can be distributed and executed in an IDL virtual machine with no need for users to purchase an IDL license.

6.2 Software Implementation for in-flight Processing of PALS Data

The software package implemented in this study aims to streamline and integrate the conventional flow of processing profiling LiDAR data, which mainly include finding ground, and segmenting profiles, and applying prediction model to convert processed PALS measurements to forest structural information either at segment or regional levels. I started to develop the software based on the existing IDL routines that had been programmed by Dr. Ross Nelson for processing PALS data, and a lot of his codes were re-used with his permission (Ross Nelson, personal communication). In the software, the basic data processing unit is a flight line. Therefore, the software in a typical mode initiates data processing modules after each flight line is done, similar to some previously developed real-time remote sensing systems that process images scene by

scene. The software was also designed to be able to start data processing once all flight lines are collected. The choice of processing mode can be made based on the extent of the study area. For a small area, the once-for-all mode is workable, but for a large area with long flight lines, the typical per-flight line mode is preferred in order to apportion the computational overhead across the flight time.

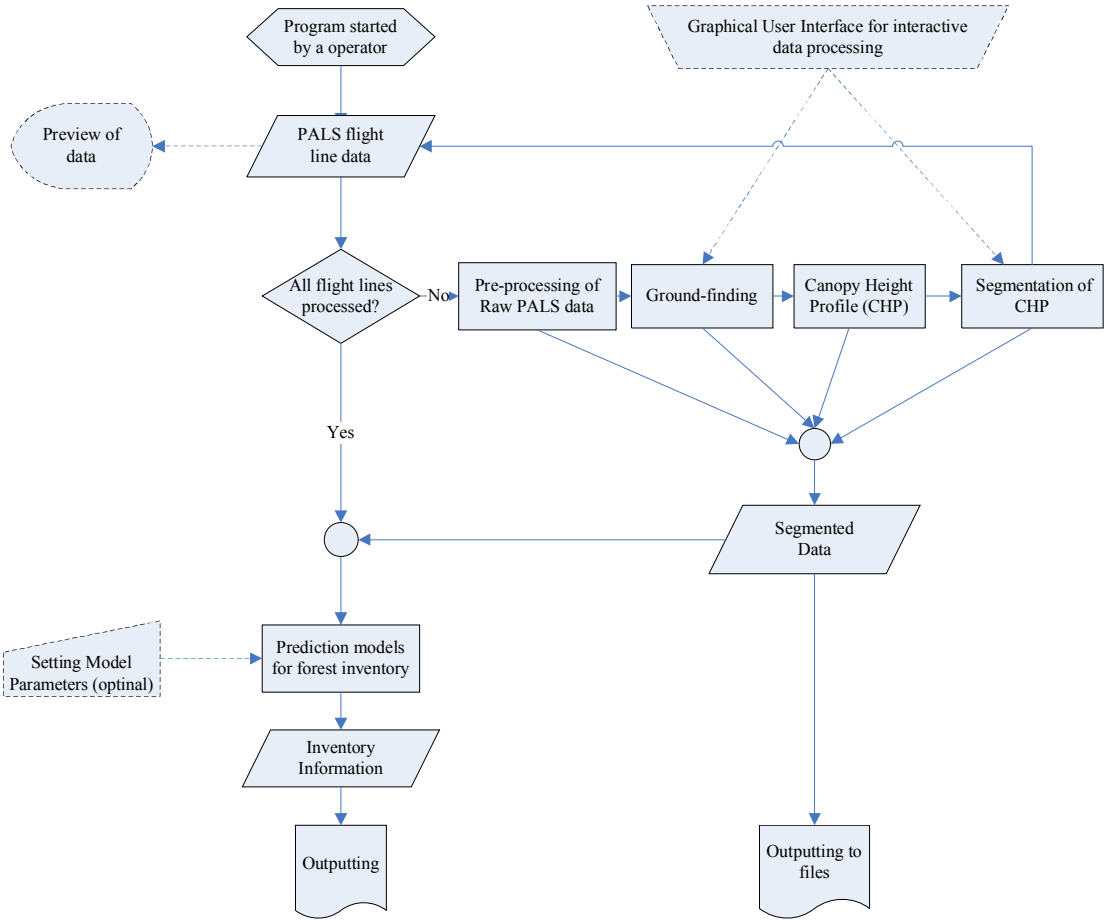


Figure 6.1.A simplified flowchart of the IDL program for on-board data processing.

The software framework for the once-for-all mode is outlined in Figure 6.1, and the per-flight line mode has a similar framework except that the ingested data are only for a single flight line. In Figure 6.1, modules outlined in dash are optional and are provided for users' conveniences. For example, if analysts believe that there exist

conspicuous errors in ground curves, they can correct ground curves interactively by adding new ground hits or eliminating misclassified hits.

6.3 Overview of the Software

Several key components of the software are described, with the aid of program snapshots if available. The PALS data used in the snapshot for illustration purposes are those collected over the intensive study area in Huntsville, east Texas.

6.3.1 PALS Raw Data Format

PALS records aircraft altitude information, such as dGPS readings, and aircraft heading and speed, every 2 seconds. Thus, raw laser data collected by PALS is organized and stored by chunks. Each chunk consists of aircraft altitude information and all laser pulses measured before the next dGPS reading. Figure 6.2 illustrates the data structure of PALS measurements. Of particular note is that in cases especially when a water surface is intercepted, reflected energies are not enough to trigger the PALS receiver so that no hits are recorded. These "no-return" pulses are indicated by a specific code, i.e., a zero value of ranging.

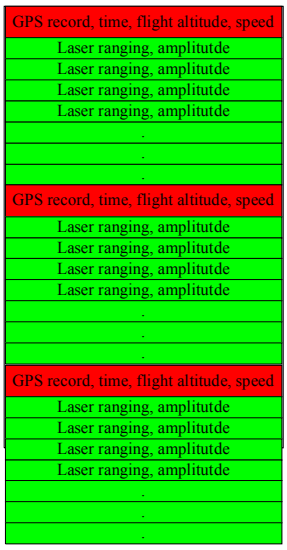


Figure 6.2. Data structure of PALS raw laser measurements.

6.3.2 Pre-processing and Preview of PALS Data

The pre-processing module interpolates 0.5 hz dGPS readings to geo-reference each laser hits, and it also identifies those pulses that have no returns. To facilitate locating flight lines, a pre-view module in the software is provided to help navigate through flight lines interactively, and an example of flight lines observed in a mission is given in Figure 6.3.

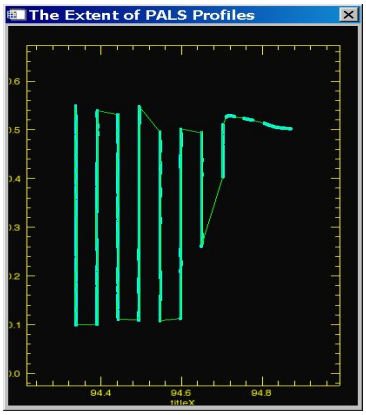


Figure 6.3. A navigation panel that shows trajectory of flight lines and also allows users to navigate through flight lines for viewing data.

6.3.3 Program Main Interface

Figure 6.4 depicts the main interface for the program. The buttons within the right-hand vertical provide links to execute each function module; the display window is used to visualize PALS raw or processed data by drawing LiDAR profiles and other relevant information, such as ground curves and strata, if available, where different symbols or colors are used to represent different information; and the menu bar on the top provides various functionalities to set up the program.

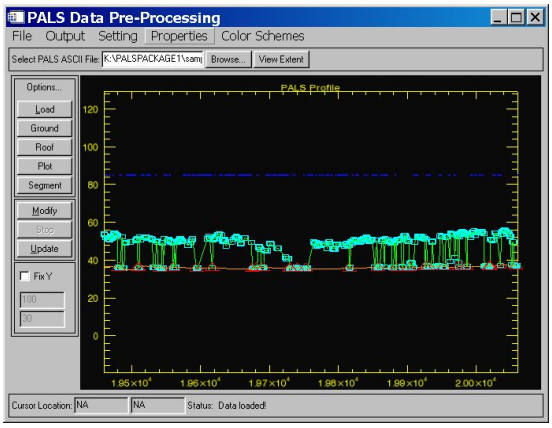


Figure 6.4. A snapshot of the main interface of the program.

6.3.4 Automatic and Interactive Ground-finding

The iterative local minimum filtering algorithm for ground-finding was implemented into the software. As discussed in Chapter II, there occasionally exist conspicuous errors in the identified terrain that cannot be rectified by the algorithm. For example, an abrupt dip or well may be identified wrongly as a ground hit, thus causing a downward erroneous ground curve, as shown in Figure 6.5a where the top rugged curve is the canopy envelop and the bottom one is the terrain with triangles indicting the identified ground hits. Figure 6.5b indicates that the program allows interactively deleting and adding ground hits, and an updated curve after correction appears more reasonable.

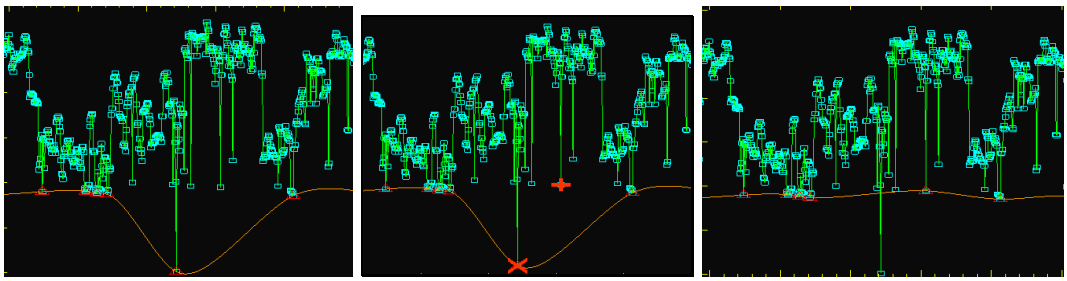


Figure 6.5. An illustration of the interactive rectification of ground curve

6.3.5 Segmentation of Canopy Height Profile

The segmentation algorithm currently implemented mainly distinguishes forest and non-forest. Roofs in the non-forest stratum will also be identified. The processed and segmented canopy height profile can be viewed or edited interactively through the main interface as shown in Figure 6.6. For later reference, each pulse of the processed and segmented data can also be saved to files according to a format illustrated in Figure 6.7.

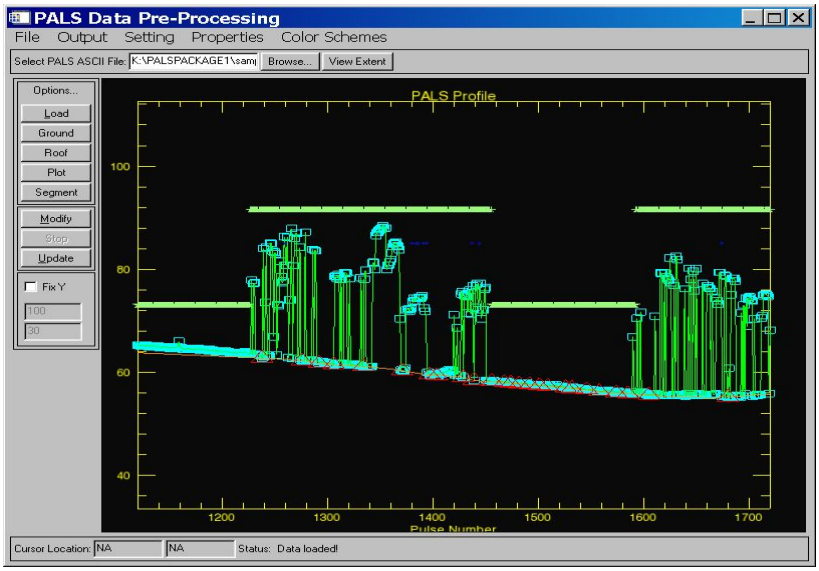


Figure 6.6. A PALS height profile that has been processed and segmented where the horizontal bars represents the resulting strata with the upper ones and lower ones being forest and non-forest, respectively.

Pulse ID	Lat, long	Altitude	heading	speed	ranging	ground	amplitude	strata
1	30.65, -95.41	375.8	10.4	205.3	257.8	258.3	20	1
2	30.65, -95.41	375.8	10.4	205.3	259.8	257.3	10	1
.....								
5770	30.67,-95.41	385.8	3	199	257.8	257.8	27	0

Figure 6.7.The format of output files for processed and segmented PALS data.

6.3.6 Using Models to Predict Forest Attributes from PALS Measurements

As a final step, the analyzed PALS data are converted into forest inventory information by using prediction models. These models are usually built based on pairs of ground-measured and coincident LiDAR measurements. However, ground-measured field data may be not available prior or during the flight such that no such models are directly available. As discussed earlier in Chapter II, generic models or those developed previously for the same study area or similar forest conditions could be used. To this end, a list of default models that were established in previous studies is maintained by the program. Users can choose one or more of these models to make prediction. Moreover, the program also allows users to build their own models by specifying model equations. To facilitate the model building, a set of the most commonly used LiDAR predictors, such as canopy height metrics and canopy density metrics, are provided to users' choices (Figure 6.8).

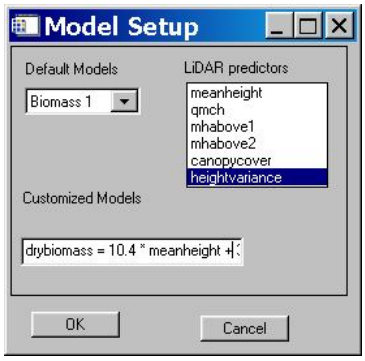


Figure 6.8. The program interface for setting up prediction models.

Models make prediction at segment levels, with either fixed- or varying- length segments. The model outputs can be exported to files. The files consist of a series of segment-level record. Each record summarizes the segment information such as starting and ending pulse IDs, segment length, and coordinates of the middle pulse, as well as the predictions from the selected models, as demonstrated in Figure 6.9.

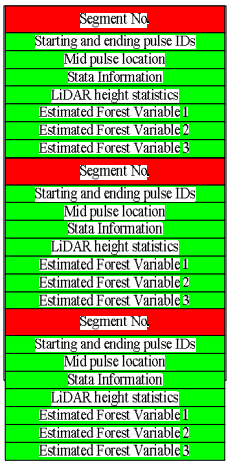


Figure 6.9. File format of outputs predicted from models

6.3.7 Parameter Setup for the Program

The program is flexible enough to allow users to set up parameters used in the algorithms if needed (Figure. 6.10). In addition, the program offers other user-friendly features for easy operation; for example, a number of shortkeys are available to provide quick operations such as panning and zooming of PALS height profiles.

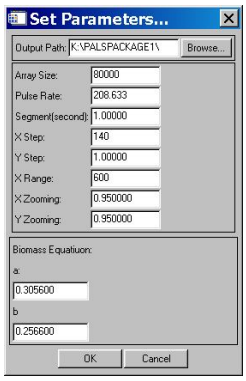


Figure 6.10. The program interface for setting up algorithm parameters.

6.4 Summary

An integrated software package was developed that incorporates the PALS data processing algorithms and LiDAR-based prediction models in order to extract forest inventory information from PALS measurements in real-time. The software requires only minimum or no human inputs so as to achieve automatic data processing. But, to allow users more flexibility, interactive modules with user-friend interfaces are also provided in case that human intervention in the data processing flow is needed. Moreover, the software, implemented in IDL, is easy to be extended if other modules are desired, e.g., when users need their own LiDAR predictors other than the default ones. As a pilot study, this research tested the software in an ad hoc manner using the existing PALS data previously collected in Huntsville, eastern Texas. It is expected that a flight will be arranged in the near future to test the software in the air.

CHAPTER VII
SUMMARY AND CONCLUSIONS

This work considers establishing a framework for an on-the-fly airborne profiling laser system in order to inventory regional forest resources in real- or near real- time. The proposed system features on-board data processing, and is able to produce instant assessments of forest resources once the airplane lands after the final fight line is flown. Such a system was developed in this work based on an existing portable airborne laser system (PALS) that has been previously assembled at NASA by Dr. Ross Nelson (Nelson et al., 2003). Key issues in automating PALS as an on-the-fly system were addressed, including the design of an archetype for the system workflow, the development of efficient and robust algorithms for automatic data processing and analysis, the development of effective regression models to predict forest biophysical parameters from LiDAR measurements, and the implementation of an integrated software package to incorporate all the above development.

The on-the-fly laser system developed in this study is of both scientific and practical significances in that it brings a reliable and affordable tool for repeatedly, quickly, and accurately assessing regional forest resources, such as forest volume, carbon stock, and biomass, at large scales such as counties, states, regions, and even continents. In the context of global environment changes, there is an arising interest in better understanding the carbon stocks of terrestrial ecosystems where forests serve as a major reservoir of carbon and play important roles in dictating the global carbon cycle. The need for spatially-explicit mapping of forest above-ground biomass over regions, continents, or the globe is resolved mostly by remote sensing. LiDAR proves to be one of the most successful remote sensing techniques for producing high-accuracy biomass map under a variety of forest conditions whereas conventional remote sensing such as optical imaging has difficulties in measuring medium to high biomass. At present, airborne scanning LiDAR, though established as an effective tool to measure canopy vertical structure, remains costly for data acquisition particularly over large areas.

Repeated measurements, which are crucial to understand changes in carbon stocks over time or monitor forest disturbances, often cannot be routinely operated with laser scanners even over local areas. On the other hand, there is a trend in the remote sensing community to build real-time remote sensing systems especially for meeting special needs of quick turn-round applications; however, in the current stage, laser scanners remains difficult to automate for airborne forestry remote sensing due to such factors as their large data volume, high costs, intensive computation, and complicated processing procedures. To these ends, the on-the-fly laser system of this study provides a viable solution to overcome all these difficulties by providing an economical means for accurately inventorying forest resources like biomass on large regions with frequent observations.

This work exploited the untouched potential of airborne laser profilers for real-time forest inventory, and therefore, documented an initial step toward developing airborne-laser-based, on-the-fly, real-time, forest inventory systems. Although the data processing algorithms cannot be directly adapted to laser scanner system, the on-the-fly airborne laser profiler developed in this study can serve as a paradigm for future efforts in building more advanced airborne laser systems such as real-time laser scanners. The algorithms currently implemented still leave room to improve. Future studies may continue investigating more sophisticated data processing algorithms for more robustness or functionalities. One possible way to improve segmentation, for example, is to examine the usefulness of the moving-window-based frequency analysis that has been used to analyze 1-D sound signals.

The model development of deriving useful LiDAR predictors for predicting forest biophysical parameters is as important as, if not more than, the algorithms development of data processing. Primary LiDAR research efforts have been focused on searching the most effective LiDAR metrics in an appropriate model form for a given forest attribute of interest, and little commonality has been found between previously reported LiDAR predictors. In this study, a new form of model, i.e., linear functional models, with canopy height distributions as LiDAR predictors, has been proposed;

moreover, the model development is justified by a theoretical treatment. The use of LiDAR-derived canopy height distribution as predictor is based on the consideration that as much information in the LiDAR measurements are preserved as possible, since most previously used predictors are simple statistics that can be extracted from canopy height distribution, but not vice versa. Although this research only looks at the use of the model for biomass, it is expected that it is equally effective for other relevant forest structural variables. Thus, it is strongly suggested that future studies examine its utility in predicting other forest attributes of interest under a variety of forest conditions.

No accepted architecture exists for developing a software package for various real-time systems because earlier studies showed that the implementation of computer programs is often task-specific. However, the integrated software package that has been implemented in IDL should be able to be transplanted to other PALS-like laser profilers, either airborne- or ground-based. In this study, the software was only tested in an ad-hoc manner using previously collected PALS data to simulate the scenario of data acquisition. It is expected that in the near future, a real flight will be scheduled to test the software and the system on air.

Continuing interests arise in employing LiDAR systems for mapping vegetation information at various scales, e.g., plot, stand, local, regional, and global. As more and more LiDAR systems, especially advanced systems equipped with scanning capabilities and high-precision navigation system, are operated and more and more LiDAR datasets, especially those with high density of laser hits and continuous coverage, become available, the profiler systems, such as the one developed in this study, may lose their positions in decades, as it first appears. I cannot admit that this conclusion is a hasty one. The original purpose of developing the system, however, is not to map forest or trees at local levels, but to provide an effective and operational tool that could be used for mapping large areas with reasonable costs.

REFERENCES

Ackermann, F. (1999). Airborne laser scanning – present status and future expectations. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(2-3), 64-67.

Affleck, D.L.R., Gregoire, T.G., & Valentine, H.T. (2005). Design unbiased estimation in line intersect sampling using segmented transects. *Environmental and Ecological Statistics*, *12*(2), 139-154.

Alerich, C.A., Klevgard, L., Liff, C., & Miles, P.D. (2004) The forest inventory and analysis database: database description and users guide, version 1.7: http://ncrs2.fs.fed.us/4801/fiadb/fiadb_documentation/FIADB_v17_122104.pdf, last accessed on June 20, 2008.

Andersen, H. E., McGaughey, R. J., & Reutebuch, S.E. (2005). Estimating forest canopy fuel parameters using LiDAR data. *Remote Sensing of Environment*, *94*, 441-449.

Aster, R., Borchers, B., & Thurber, C. (2004). *Parameter Estimation and Inverse Problems*. Academic Press, New York, 230 pp.

Axelsson, P. (1999). Processing of laser scanner data – algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(2-3), 138-147.

Barilotti, A., Turco, S., Alberti, G. (2006). LAI determination in forestry ecosystems by LiDAR data analysis. *Workshop on 3D Remote Sensing in Forestry*, 14-15/02/2006, BOKU Vienna.

Bauer, H. L. (1943). The statistical analysis of chaparral and other plant communities by means of transect samples. *Ecology*, *24*, 45-60.

Birky, A.K. (2001). NDVI and a simple model of deciduous forest seasonal dynamics, *Ecological Modelling*, *143*, 43–58.

Brandtberg, T., Warner, T.A., Landenberger, R.E. & McGraw, J.B. (2003). Detection and analysis of individual leaf-off tree crowns in small footprint, high sampling density LiDAR data from the eastern deciduous forest in North America. *Remote Sensing of Environment*, *85*, 290–303.

Butler, S. & McDonald, L.L. (1983). Unbiased systematic sampling plans for the line intercept method. *Journal of Range Management*, *36*, 463–468.

Canfield, R. H. (1941). Application of the line interception method in sampling range vegetation. *Journal of Forestry*, *39*, 388-394.

Carter, G.A. (1991). Primary and secondary effects of water content of the spectral reflectance of leaves. *American Journal of Botany*, *78*, 916–924.

Chan, D., Higuchi, K., Shashkov, A., Liu, J., Yuen, C. W., Chen, J., Worthy, D. (2003). On the CO2 exchange between the atmosphere and the biosphere: the role of synoptic and mesoscale processes, *Tellus*, *Series B*, *56B*(3), 194-212.

Chang, C.-I., Ren, H., Chiang, S.-S. (2001). Real-time processing algorithms for target detection and classification in hyperspectral imagery. *IEEE Transaction on Geosciences and Remote Sensing*, *39*(4), 760-768.

Chen, J. M., & Black, T. A. (1992). Defining leaf area index for non-flat leaves. *Plant, Cell and Environment*, *15*, 421– 429.

Chen, Q., Baldocchi, D.D., Gong, P., & Kelly, M. (2006). Isolating individual trees in a savanna woodland using small footprint LIDAR data, *Photogrammetric Engineering and Remote Sensing*, *72*(8), 923-932.

Chen, Q., Gong, P., Baldocchi, D.D., & Tian, Y. (2007). Estimating basal area and stem volume for individual trees from LIDAR data. *Photogrammetric Engineering and Remote Sensing*, *73*(12), 1355-1365

Cohen, W.B., Maiersperger, T. K., Turner, D. P., Ritts, W. D., Pflugmacher, D., Kennedy, R. E., Kirschbaum, A., Running, S. W., Costa, M., & Gower, S. T. (2006). MODIS land cover and LAI collection 4 product quality across nine sites in the Western Hemisphere. *IEEE Transaction on Geosciences and Remote Sensing*, *44*, 1843.

Cohen, W.B., & Goward, S.N. (2004). Landsat's role in ecological applications of remote sensing. *BioScience*, *54*, 535–545.

Cohen, W.B., Maiersperger, T.K. , Gower, S.T., & Turner, D.P. (2003). An improved strategy for regression of biophysical variables and landsat etm+ data. *Remote Sensing of Environment, 84*, 561–571.

Coifman, R.R., Donoho, D.L. (1995). Translation-invariant denoising. In Antoniadis, A. and Oppenheim, G. (Eds), *Wavelet and Statistics*, *Lecture Notes in Statistics*, Springer, Berlin, 1995, 125-150.

Crofton, M. W. (1868). On the theory of local probability, applied to straight lines drawn at random in a plane. *Philosophical Transactions of the Royal Society of London*, *158*, 181-199.

Derrien, M, Farki, B., Harang, L., LeGleau, H., Noyalet, A., Pochic, D., Sairouni, A. (1993). Automatic cloud detection applied to NOAA-11/AVHRR imagery. *Remote Sensing of Environment*, *46*(3), 246-267.

Drake, J. B., Dubayah, R. O., Clark, D. B., Knox, R. G., Blair, J. B., Hofton, M. A., Chazdon, R. L., Weishampel, J. F., & Prince, S. D. (2002). Estimation of tropical forest structural characteristics using large-footprint LiDAR. *Remote Sensing of Environment*, *79,* 305– 319.

Durbha, S.S., King, R.L. & Younan, N.H. (2007). Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment*, *107*, 348-361.

Eklundh, L., Harrie L., & Kuusk, A. (2001). Investigating relationships between Landsat ETM+ sensor data and leaf area index in a boreal conifer forest, *Remote Sensing of Environment*, *78*, 239–251.

Elmqvist, M., Jungert, E. , Lantz, F., Persson, A., & Söderman U. (2001). Terrain modeling and analysis using laser scanning scanner. *International Archives of Photogrammetry and Remote Sensing*, XXXIV-3/W4, 2001, 219-226.

EOS website: http://eospso.gsfc.nasa.gov/, last accessed on March 15, 2006

Gutierrez, R., Neuenschwander, A., & Crawford, M.M. (2005). Development of laser waveform digitization for airborne LiDAR topographic mapping instrumentation. *Proceedings of the IEEE IGARSS Conference*, *2*, 1154-1157.

Holmgren, J. (2004). Prediction of tree height, basal area and stem volume using airborne laser scanning. *Scandinavian Journal of Forest Research*, *19*, 543–553.

Honda K., & Nagai, M. (2002). Real-time volcano activity mapping using ground-based digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *57*, 144–153.

Hopkinson, C., & Chasmer, L. (2007). Modelling canopy gap fraction from LiDAR intensity. In Rönnholm, P., Hyyppä, H., Hyyppä, J. (Eds). *Proceedings of the ISPRS Working Group "Laser Scanning 2007 and SilviLaser 2007"*, ISPRS Volume XXXVI, Part3/W52. Espoo, September 12-14, 2007, Finland, 190-194.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Xiaosu, D., Maskell, K., & Johnson, C. A. (Eds). (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK7 Cambridge University Press, 944 pp.

Houghton, R.A. (2005). Aboveground forest biomass and the global carbon balance. *Global Change Biology*, *11*, 945–958.

Houghton, R.A. (2007). Balancing the global carbon budget. *Annual Review of Earth and Planetary Sciences*, *35*(1), 313

Hsu, SM., Burke, H., Griffin, M. (2001). Adaptive HSI data processing for near-real-time analysis and spectral recovery. *Proceedings of the Tenth JPL Airborne Earth Science Workshop*, NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, 229-238.

Hugershoff, R. (1939). Die Bildmessung und ihre forstlichen Anwendungen. *Der Deutsche Forstwirt*, *21*(50), 612-615.

Hyyppä, J.M., & Hallikainen MT. (1996). Applicability of airborne profiling Radar to forest inventory. *Remote Sensing of Environment*, *57*, 39-57.

Jenkins, J.C., Chojnacky, D.C., Heath, L.S., & Birdsey, R.A. (2003). National-scale biomass estimators for United States tree species. *Forest Science*, *49*, 12–35.

Jonckheere, I., Fleck, S., Nackaerts, K., Muys, B., Coppin, P., Weiss, M., & Baret, F. (2004). Review of methods for in situ Leaf Area Index determination part I. Theories, sensors, and hemispherical photography, *Agricultural and Forest Meteorology*, *121*, 19–35.

Kaiser, L. (1983). Unbiased estimation in line-intercept sampling. *Biometrics*, *39*, 965-976.

Keane, R.E., Reinhardt, E.D., Scott, J., Gray, K., & Reardon, J. (2005). Estimating forest canopy bulk density using six indirect methods. *Canadian Journal of Forestry Research*, *35*, 724–739.

Keller, M., Palace, M., Asner, G.P., Pereira, R., & Silva, J.N.M. (2004). Coarse woody debris in undisturbed and logged forests in the eastern Brazilian Amazon. *Global Change Biology*, *10*, 784–795.

Konare, D., Pierre, S., Weng, J.Y., & Morand, E. (2003). Real-time image processing for remote sensing. In: *Proceedings - Toward a Caring and Humane Technology*, *Vols 1-3*, CCECE 2003: Canadian Conference on Electrical and Computer Engineering, 699-702.

Kotchenova, S.Y., Shabanov, N.V., Knyazikhin, Y., Davis, A.B., Dubayah, R. and Myneni, R.B. (2003). Modeling LiDAR waveforms with time-dependent stochastic

radiative transfer theory for remote estimations of forest structure. *Journal of Geophysical Research*, *108*(D15), p. 4484

Kraus, K., & Pfeifer, N. (1998). Determination of terrain models in wooded areas with airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *53*, 193-203.

Lavender, S.J. & Groom, S.B. (1999). The seaWiFS automatic data processing system (SeaAPS). *International Journal of Remote Sensing*, *20*, 1051–1056.

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A. & Harding, D., (1999). LiDAR remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. *Remote Sensing of Environment*, *70*, 339–361.

Lefsky, M.A., Harding, D., Cohen, W.B., Parker, G., & Shugart, H.H. (1999). Surface LiDAR remote sensing of basal area biomass in deciduous forests of eastern Maryland, USA. *Remote Sensing of Environment*, *67*, 83-98.

Lefsky, M.A., Cohen, W.B., Parker, G.G., & Harding, D.J. (2002). LiDAR remote sensing for ecosystem studies, *Bioscience*, *52*(1), 19-30

Lefsky, M.A., Cohen, W.B. , Acker, S.A., Spies, T.A., Parker, G.G., & Harding, D. (1997). LiDAR remote sensing of forest canopy structure and related biophysical parameters at the H.J. Andrews experimental forest, Oregon, USA. In *Natural Resources Management Using Remote Sensing and GIS*. Greer, J.D., Ed., ASPRS, Washington, D.C., 79-91.

Li, F., Zhang, L., & Davis, C.J. (2002). Modeling the joint distribution of tree diameters and heights by bivariate generalized beta distribution. *Forest Science*, *48*(1), 47–58.

Li, X., & Strahler, A.H. (1992). Geometric-optical bidirectional reflectance modeling of the discrete crown vegetation canopy: Effect of crown shape and mutual shadowing, *IEEE Transaction on Geosciences and Remote Sensing*, *30*, 276–292.

Li, Z., Nadon, S., & Cihlar, J. (2000). Satellite-based detection of Canadian boreal forest fires: development and application of the algorithm. *International Journal of Remote Sensing*, *21*(16), 3057-3070.

Liang, S. (2004). *Quantitative Remote Sensing of Land Surfaces*. New York: John Wiley.

Lienert, B. R., Porter, J. N., & Sharma, S. K. (1999). Real time analysis and display of scanning lidar scattering data. *Marine Geodesy*, *22*, 259–265.

Lim, K., Treitz, P., Baldwin, K., Morrison, I. & Green, J. (2003). LiDAR remote sensing of biophysical properties of tolerant northern hardwood forests. *Canadian Journal of Remote Sensing*, *29*, 648–678.

Lim, K.S., & Treitz, P.M. (2004). Estimation of above ground forest biomass from airborne discrete return laser scanner data using canopy-based quantile estimators, *Scandinavian Journal of Forest Research*, *19*(6), 558–570.

Lindeberg, J. W. (1926): Zur Theorie Derr Linientaxierung. *Acta Forestalia Fennica*, *31*(6), 3 –9.

MacArthur, R. H., & Horn, H. S. (1969). Foliage profiles by vertical measurements. *Ecology*, *50,* 802-804.

MacLean, G.A., & Krabill, W.B. (1986). Gross-merchantable timber volume estimation using an airborne LIDAR system. *Canadian Journal of Remote Sensing*, *12*(1), 7-18.

Maeda, Y., Tsuzuki, H., Nelson, R., & Sweda, T. (2008). Land-cover classification of Ehime Prefecture, Japan using airborne laser altimetry. *Journal of Forest Planning (Japanese Society of Forest Planning)*, *13*, 245-248.

Maltamo, M, Eerikäinen, K., Packalén, P., & Hyyppä, J. (2006). Estimation of stem volume using laser scanning-based canopy height metrics. *Forestry*, *79*(13), 217-229.

Maltamo, M., Eerikaeinen, K., Pitkaenen, J., Hyyppae, J., Vehmas, M. (2004). Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sensing of Environment*, *90*(3), 319-330.

Maltamo, M., Kangas, A., Uuttera, J., Torniainen, T., & Saramäki, J. (2000). Comparison of percentile based predicted methods and Weibull distribution in describing diameter distribution of heterogeneous Scots pine stands. *Forest Ecology and Management*, *133*, 263-274.

Marceau, D.J. (1999). The scale issue in social and natural sciences. *Canadian Journal of Remote Sensing*, *25*, 347-356.

Marceau, D.J., Howarth, P.J., & Gratton, D.J. (1994). Remote sensing and the measurement of geographical entities in a forested environment. 1: The scale and spatial aggregation problem, *Remote Sensing of Environment*, *49*(2), 93-104.

Masayuki, M., & Hidefumi, N. (2002). An advanced airborne real-time SAR processing system. *Proceedings of IGARSS 2002*, CA, 856-858.

McCombs, J.W., Roberts, S.D., Evans, D.L. (2003). Influence of fusing LiDAR and multispectral imagery on remotely sensed estimates of stand density and mean tree height in a managed Loblolly pine plantation. *Forest Science*, *49*(3) 457-466.

McIntyre, G. A. (1953). Estimation of plant density using line transects. *Journal of Ecology*, *41*, 319-330.

Means, J.E., Acker, S.A., Harding, D.J., Blair, J.B., Lefsky, M.A., Cohen,W.B., Harmon, M., & McKee, W.A. (1999). Use of large-footprint scanning airborne LiDAR to estimate forest stand characteristics in the western Cascades of Oregon. *Remote Sensing of Environment*, *67*, 298–308.

Millette, T. L., & Hayward, C. D. (2008). Detailed forest stand metrics taken from AIMS-1 sensor data:http://www.mtholyoke.edu/dept/earth/facilities/Millette_b.pdf, last accessed on Feb. 16, 2008.

Morsdorf, F., Kötz, B., Meier, E., Itten, K. I., & Allgower, B. (2006). Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction. *Remote Sensing of Environment*, *104*, 50-61.

Mutlu, M., Popescu, S.C., Stripling, C., & Spencer, T. (2008). Assessing surface fuel models using LiDAR and multispectral data fusion. *Remote Sensing of Environment*, *112*, 274-285.

Myneni, R. B., Ramakrishna, R., & Running, S. W. (1997). Estimation of global leaf area index and absorbed PAR using radiative transfer models. *IEEE Transactions on Geoscience and Remote Sensing*, *35*, 1380–1393.

Næsset, E. & Bjerknes, K.-O. (2001). Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. *Remote Sensing of Environment*, *78*, 328–340.

Næsset, E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*, *61*, 246– 253.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, *80*, 88-99.

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

Nason, G.P., & Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. In Antoniadis, A. and Oppenheim, G. (Eds), *Wavelet and Statistics*, *Lecture Notes in Statistics*, Springer, Berlin, 281-300.

Nelson, R. (1997). Modeling forest canopy heights: The effects of canopy shape. *Remote Sensing of Environment*, *60*, 327-334.

Nelson, R., Short, A., & Valenti, M. (2004). Measuring biomass and carbon in Delaware using an airborne profiling LiDAR. *Scandinavian Jour. of Forest Research*, *19*, 500-511. [Erratum. 2005, 3: 283-284.]

Nelson, R., Keller, C., & Ratnaswamy, R. (2005). Locating and estimating the extent of Delmarva fox squirrel habitat using an airborne LiDAR profiler. *Remote Sensing of Environment*, *96*(3-4), 292-301.

Nelson, R., Parker, G., & Horn, M. (2003a). A portable airborne laser system for forest inventory. *Photogrammetric Engineering and Remote Sensing*, *69*, 267–273.

Nelson, R., Valenti, M. A., Short, A., & Keller, C., 2003b. A multiple resource inventory of Delaware using airborne laser data. *Bioscience*, *53*, 981– 992.

Nelson, R., Oderwald, R.G., & Gregoire, T.G. (1997). Separating the ground and airborne sampling phases to estimate tropical forest basal area, volume, and biomass. *Remote Sensing of Environment*, *60*, 311-326.

Nelson, R., Swift, R., & Krabill, W. (1988a). Using airborne lasers to estimate forest canopy and standcharacteristics. *Journal of Forestry*, *86*, 31-38.

Nelson, R., Krabill, W., & Tonelli, J. (1988b). Estimating forest biomass and volume using airborne laser data. *Remote Sensing of Environment*, *24*, 247-267.

Nelson, R., Krabill, W., & Maclean, G.A. (1984). Determining forest canopy characteristics using airborne laser data. *Remote Sensing of Environmen*t, *15*, 201–212.

Nilson, T. (1999). Inversion of gap frequency data in forest stands. *Agricultural and Forest Meteorology*, *98-99*, 437-448.

Omasa, K., Qiu, G. Y., Watanuki, K., Yoshimi, K., & Akiyama, Y. (2003). Accurate estimation of forest carbon stocks by 3-D remote sensing of individual trees. *Environmental Science and Technology*, *37*, 1198–1201.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth Sciences. *Science*, *263*, 641–646.

Parresol, B.R. (1999). Assessing tree and stand biomass: a review with examples and critical comparisons. *Forest Science*, *45*(4), 573-593

Patenaude, G., Hill, R.A., & Milne, R. (2004). Quantifying forest above ground carbon content using LiDAR remote sensing. *Remote Sensing of Environment*, *93*, 368–380.

Pearson, R, Grace, J, & May, G. (1994). Real-time airborne agricultural monitoring. *Remote Sensing of Environment*, *49*(3), 304-310.

Petzold, B., Reiss P., & Stössel W. (1999). Laser scanning - surveying and mapping agencies are using a new technique for the derivation of digital terrain models. *ISPRS Journal of Photogrammetry & Remote Sensing*, *54*(2-3), 95-104.

Plaza, (2008). http://www.umbc.edu/rssipl/people/aplaza/Papers/Journals/2008. JRTIP.Cfp.pdf, last accessed on Apr. 7, 2008

Popescu, S.C. & Wynne, R.H. (2004). Seeing the trees in the forest: using LiDAR and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing*, *70*(5), 589-604.

Popescu, S. C., Wynne, R. H., & Nelson, R.H., 2002. Estimating plot-level tree heights with LIDAR: Local filtering with a canopy-height based variable window size. *Computers and Electronics in Agriculture*, *37*(1-3), 71−95.

Popescu, S. C.,Wynne, R. H.,& Nelson, R. H., 2003. Measuring individual tree crown diameter with LIDAR and assessing its influence on estimating forest volume and biomass. *Canadian Journal of Remote Sensing*, *29*(5), 564−577.

Popescu, S.C. & Zhao, K.G., 2008. A voxel-based LiDAR method for estimating crown base height for deciduous and pine trees. *Remote Sensing of Environment*, *112*(3), 767-781.

Popescu, S.C. (2002). *Estimating Plot-Level Forest Biophysical Parameters Using Small-Footprint Airborne LiDAR Measurements*. Ph. D. Dissertation, Virginia Tech University, Blacksburg.

Popescu, S.C. (2007). Estimating biomass of individual pine trees using airborne LiDAR, *Biomass and Bioenergy*, *31*, 646–655.

Ramsey, M.S., Dehn, J., Wessels, R., Byrnes, J., Duda, K., Maldonado, L., & et al. (2004). *The ASTER emergency scheduling system: A new project linking near-real-time satellite monitoring of disasters to the acquisition of high-resolution remote sensing data*. American Geophysical Union Fall Meeting, San Francisco, USA, 2004 (abs. SF23A-0026).

Riaño, D., Valladares, F., Condés, S., & Chuvieco, E. (2004). Estimation of leaf area index and covered ground from airborne laser scanner (Lidar) in two contrasting forests. *Agricultural and Forest Meteorology*, *124*(3-4), 269−275.

Ritchie, J.C., Everitt, J.H., Escobar, D.E., Jackson, T.J., & Davis, M.R. (1992). Airborne laser measurements of rangeland canopy cover and distribution. *Journal of Range Management*, *45*, 189–193.

Ritchie, J.C. (1995). Airborne laser altimeter measurements of landscape topography. *Remote Sensing of Environment*, *53*, 91-96.

Ross, J.K. (1981). *The Radiation Regime and Architecture of Plant Stands*. Dr W. Junk Publishers, The Hague, Boston, London.

Sandau, R., Braunecker, B., Driescher, H., Eckardt, A., Hilbbert, S., Hutton, J., Kirchhofer, W., Lithopoulos, E., Reulke, R., & Wicki, S. (2000). Design principles of the LH Systems ADS40 airborne digital sensor. *International Archives of Photogrammetry and Remote Sensing*, *33*(B1), 258-265.

Saura, S., & Martínez-Millán, J. (2000). Landscape patterns simulation with a modified random clusters method. *Landscape Ecology*, *15*(7), 661-678.

Schreier, H., Lougheed, J., Tucker, C., & Leckie, D. (1985). Automated measurements of terrain reflection and height variations using and airborne infrared laser system. *International Journal of Remote Sensing*, *6*(1), 101-113.

Shabanov, N.V., Kotchenova, S., Huang, D., Yang, W. , Tan, B., Knyazikhin Y., & et al. (2005). Analysis and optimization of the MODIS leaf area index algorithm retrievals over broadleaf forests, *IEEE Transactions on Geoscience and Remote Sensing* ,*43* (8), 1855–1865.

Shutler, J. D., Smyth, T. J., Land, P. E., Groom, S.B., (2005). A near-real time automatic MODIS data processing system. *International Journal of Remote Sensing*, *26*, 1049-1055.

Skowronski, N., Clark, K., Nelson, R., Hom, J., & Patterson, M. (2007). Remotely sensed measurements of forest structure and fuel loads in the Pinelands of New Jersey. *Remote Sensing of Environment*, *108*(2), 123-129.

Strahler, A. H., Woodcock, C. E., & Smith, J. A. (1986). On the nature of models in remote sensing. *Remote Sensing of Environment, 20*, 121-139.

Sun, G, Ranson, K.J., Kimes, D.S., Blair, J.B., Kovacs, K. 2008. Forest vertical structure from GLAS: an evaluation using LVIS and SRTM data, *Remote Sensing of Environment*, *112*, 107–117.

Sun, G., & Ranson, K.J. (2000). Modeling LiDAR returns from forest canopies. *IEEE Transactions on Geoscience and Remote Sensing*, *38*(6), 2617–2626.

Suter, M., Nuesch, D., 1995. Automated generation of visual simulation databases using remote sensing and GIS. *IEEE Visualization '95*, Atlanta GA, pp. 86-93.

Tehranian,S., Zhao, Y., Harvey, T., Swaroop, A., & Mckenzie, K. (2006). A robust framework for real-time distributed processing of satellite data, *Journal of Parallel and Distributed Computing, 66*, 403–418.

Tewari, V.P., & Gadow, K.V. (1999). Modelling the relationship between tree diameters and heights using SBB distribution. *Forest Ecology and Management*, *119*, 171–176.

Tian, Y., Woodcock, C.E., Wang, Y. , Privette, J.L., Shabanov, N.V., & Zhou, L., & et al. (2002). Multiscale analysis and validation of the MODIS LAI product. I. Uncertainty assessment, *Remote Sensing of Environment*, *83*, 414–430.

Tickle, P., Witte, C., Danaher, T., Jones, K. (1998). The application of large-scale video and laser altimetry to forest inventory. In *Proceedings, 9th Australiasian Remote Sensing and Photogrammetry Confere*nce, Sydney, Australia, 9–13.

Toth, C.K., & Grejner Brzezinska, D.A. (2002). Near real time road centerline extraction. In *Proceedings*: *ISPRS. Photogrammet. Comput. Vision*, Graz, Austria, September, 2002, 9-13.

Turner, D.P., Ollinger, S.V., & Kimball, J.S. (2004). Integrating remote sensing and ecosystem process models for landscape- to regional-scale analysis of the carbon cycle. *BioScience*, *54*, 573–584.

Turner, M.G., Dale, V.H., & Gardner, R.H. (1989). Predicting across scales: Theory development and testing, *Landscape Ecology*, *3*, 245-252.

van Aardt, J. A.N., Wynne, R. H., & Oderwald, R.G. (2006). Forest volume and biomass Estimation using small-footprint LiDAR-distributional parameters on a per-segment basis. *Forest Science, 52*(6), 636-649.

van Gardingen, P.R., Jackson, G.E., Hernandez-Daumas, S., Russell, G., & Sharp, L. (1999). Leaf area index estimates obtained for clumped canopies using hemispherical photography. *Agricultural and Forest Meteorology*. *94*, 243-257.

Van Wagner, C.E. (1961). The line intersect method in forest fuel sampling. *Forest Science, 1*4(1), 20-26.

Wang, Q., J. Tenhunen , N. Q. Dinh , M. Reichstein , D. Otieno , A. Granier & Pilegarrd, K. (2005). Evaluation of seasonal variation of MODIS derived leaf area index at two European deciduous broadleaf forest sites. *Remote Sensing of Environment*, *96*, 475-484.

Wang, Y., Woodcock, C. E., Buermann, W., Stenberg, P., Voipio, P., Smolander, H., et al. (2004). Evaluation of the MODIS LAI algorithm at a coniferous forest site in Finland. *Remote Sensing of Environment*, 91, 114–127.

Warren, W. G. & P. F. Olsen. (1964). A line intersect technique for assessing logging waste. *Forest Science*, *10*, 267-276.

Wehr, A., & Lohr, U. (1999). Airborne laser scanning - an introduction and overview. *ISPRS Journal of Photogrammetry & Remote Sensing*, *54*(2-3), 68-82.

Woodcock, C. E. & Strahler, A.H., (1987). The factor of scale in temote sensing. *Remote Sensing of Environment*, *21*, 311-332.

Wu, S. S. C., Hutton, J., Kletzli, B., Noto, H. M., Wang, S. S.,& Wang, Z. H. (2004). A teal-time photogrammetric mapping system. In *Proceedings of the ISPRS working group "Geo-Imagery Bridging Continents"*, XXth ISPRS Congress, 12-23 July 2004, Istanbul, Turkey, p. 61.

Wulder, M. (1998). Optical remote-sensing techniques for the assessment of forest inventory and biophysical parameters. *Progress in Physical Geography*, *22*, 449–476.

Young, B., Evans, D.L., & Parker, R.C. (2000). Methods for comparison of LiDAR and field measurements of loblolly pine. In *Proceedings*: *Second International Conference on Geospatial Information in Agriculture and Foretry*, Lake Buena Vista, Florida, 10–12 January 2000: I-193–199.

Yu, X., Hyypp, J., Kaartinen, H., & Maltamo, M. (2004). Automatic detection of harvested trees and determination of forest growth using airborne laser scanning. *Remote Sensing of Environment*, *90*, 451−462.

Zavitkovski, J. (1976). Ground vegetation biomass, production, and efficiency of energy utilization in some northern Wisconsin forest ecosystems. *Ecology*, *57*, 694–706.

Zhao, K., & Popescu, S.C. (2007). Hierarchical watershed segmentation of canopy height model for multi-scale forest inventory. In Rönnholm, P., Hyyppä, H., Hyyppä, J. (eds). *Proceedings of the ISPRS Working Group "Laser Scanning 2007 and SilviLaser 2007"*, ISPRS Volume XXXVI, Part3/W52. Espoo, September 12-14, 2007, Finland, 436-442.

Zhao, K., Popescu, S.C., & Zhang, X. (2008). Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. *Photogrammetric Engineering & Remote Sensing* (in press).

Zhao, K., S. C. Popescu, & R. F. Nelson, 2008. Quantifying the uncertainty for the line-intercept sampling estimators of canopy cover, *Journal of Forest Planning (Japanese Society of Forest Planning)*, *13*, 195-205.

Zimble, D.A., Evans D.L., Carlson G.C., Parker R.C., Grado, S.C., & Gerard P.D. (2003). Characterizing vertical forest structure using small-footprint airborne LiDAR. *Remote Sensing of Environment*, *87*, 171-182.

VITA

Kaiguang Zhao

Spatial sciences lab,
Dept. of Ecosystem Science and Management,
Texas A&M University, College Station, TX 77843
Tel: 1-979-458-0742 (O)
Email: lidar.rs@gmail.com

Education

Ph.D. in Forestry with emphasis on forestry remote sensing, Texas A&M University, Aug. 2008
GPA: 4.0

M.S. in Cartography and Geography Information Systems, Beijing Normal University, China, Aug. 2004

B.S. in Physics, Beijing Normal University, China, Aug. 2001

Awards & Honors

- Who's Who Among Students in American Universities and Colleges, 2008
- Outstanding Ph.D. Student of the Year, ESSM, Texas A&M University, 2008
- The Third Place in Student Paper Competition, SWAAG/Mid-south ASPRS Conference, 2007
- ASPRS Leica-Geosystems Internship Award, the American Society for Photogrammetry & Remote Sensing, Aug. 2007
- The Honor Society of Phi Kappa Phi, 2007-present.
- International Education Fee Scholarship, Texas A&M University, 2007

Refereed Publications

M. Mutlu, S.Popescu, & K. Zhao, (in press). Sensitivity Analysis Of Fire Behavior Modeling With Lidar-Derived Surface Fuel Maps, Forest Ecology and Management.

K. Zhao, S. Popescu, & X. Zhang, (in press). Bayesian Leaning with Gaussian Processes for Supervised Classification of Hyperspectral Data, Photogrammetry Engineering & Remote Sensing.

S. Popescu & K. Zhao, 2008. A Voxel-based Lidar Method for Estimating Crown Base Height for Deciduous and Pine trees, Remote sensing of Environment: 112(3): 767-781.

K. Zhao, S. C. Popescu, & R. F. Nelson, 2008. Quantifying the Uncertainty for the Line-intercept Sampling Estimators of Canopy Cover, Journal of Forest Planning, 13: 195-205

K. Zhao & S. Popescu, 2007. Hierarchical Watershed Segmentation of Canopy Height Model for Multi-scale Forest Inventory, ISPRS Volume XXXVI, Part 3 / W52, 436-442.