A Light Detection and Ranging (Lidar) study of the Sierra Nevada with geo-spatial applications

By

Gary M. Phelps II

B.S, Computer Science and Engineering (2009)
University of California, Merced

Submitted to the Program of Environmental Systems and School of Engineering
In Partial Fulfillment of the Requirement for the Degree of

Master of Science
in Environmental Systems

at the

University of California, Merced

May 2011

Author _____
Gary M. Phelps II
Graduate Student, UC Merced

Certified by _____
Qinghua Guo
Committee Chair
Assistant Professor, UC Merced

Certified by _____
Thomas Harmon
Professor, UC Merced

Certified by _____
Shawn Newsam
Assistant Professor, UC Merced

A Light Detection and Ranging (Lidar) study of the Sierra Nevada with geo-spatial applications

By
Gary M. Phelps II

Submitted to the Program of Environmental Systems and School of Engineering of May 2nd, 2011 in partial fulfillment for the Degree of Master of Science in Environmental Systems

Abstract

Light Detection and Ranging (lidar) has been used widely for the remote sensing of multiple parameters from earth's surface. Lidar systems are used to measure light scattered to find and or range a specific target using laser pulses and radio waves by measuring the time delay between transmission of a pulse and detection of reflected signal. Lidar has proven to be a promising technology for estimating forest biophysical parameters, but due to high-cost of flights, computer processing times, hard drive storage limitations, lidar flights are not numerous and difficult to process at high-resolutions. Discreet return lidar (three dimensional point cloud data) is used for a variety of applications including: urban planning, forest management, wildlife habitat analysis, and forest biomass estimations. This study aims to provide a framework in generating lidar-derived product such as Digital Elevation Models (DEMs), Digital Surface Models (DSMs) and lidar-derived biomass estimates for a study area in the Sierra Nevada. This study also provides an open-source framework for storing and sharing spatial data using an online web-content management system. Results include USGS and lidar-derived DEM error, generating DSMs across a variety of platforms including point-density reduction, interpolation methods and resolutions, as well as a comparison of estimating biomass using individual tree extraction from lidar and a multivariate point cloud regression approach using ground-truthed plot data. The web-based software in this study is used to store and share data amongst a variety of teams and persons including the public, the Sierra Nevada Adaptive Management Project, National Critical Zone Observatories and other research teams associated with UC Merced.

Acknowledgments

First of all, I would like to thank my thesis supervisor, Assistant Professor Qinghua Guo, for giving me the opportunity to become involved in the graduate studies program at UC Merced and for his guidance and financial support over the past few years. His encouragement and advice were essential for the accomplishment of my thesis.

I am also grateful to Professor Tom Harmon for providing useful information and edits on many papers in this thesis and the support over the past year as well as the opportunity to become involved in projects such as Mar-Net and Hyperspectral Remote Sensing with his lab members.

I'd like to thank the Sierra Nevada Adaptive Management Project for their support along with the collaboration of the National Critical Zone Observatory for support on developing the digital library and the chance to present this development in front of CZO PIs at the University of Colorado at Boulder.

I would like to thank my professors, the staff, and the colleagues in Bales, Harmon and Guo Lab for providing information and many useful suggestions and help for my thesis including: Otto Alvarez, Jacob Flanagan, Wenkai Li, Hong Yu, Andrew Zumkehr and Gesha Uminskiy.

Finally I'd like to thank my mom, dad and my girlfriend Shaina for the love and support through my undergraduate and graduate years here at UC Merced.

Table of Contents

List of Figures and Tables

Chapter 1 – Introduction to Lidar data

Light Detection and Ranging (lidar) is a prominent remote sensing technology that provides 3-dimensional point cloud data that are useful for a variety of geographic and spatial applications. Small footprint, discrete return airborne lidar is essentially a laser scanner that emits discrete pulses to an object and allows for simultaneous mapping of ground, vegetation, buildings and other features. Lidar has been applied across a variety of disciplines in studies pertaining to hydrological modeling, flood prediction, canopy height and biomass estimates for forests (Toyra, *et al.* 2005; Kato *et al.* 2009; Popescu *et al.* 2004). In areas of dense vegetation or canopy cover, only a small portion of lidar pulses will penetrate the canopy; most reflect off the top and within the vegetation canopy. The laser pulses penetrating to the ground, classified as "ground-hits," are important because they enable accurate determination of ground elevations or digital elevation models (DEMs). Pulses that reflect off the top and from within the vegetative canopy are used to generate digital surface models (DSMs). *Figure 1* below displays the raw lidar point cloud viewed in the System for Geoscientific Analyses (SAGA) application.

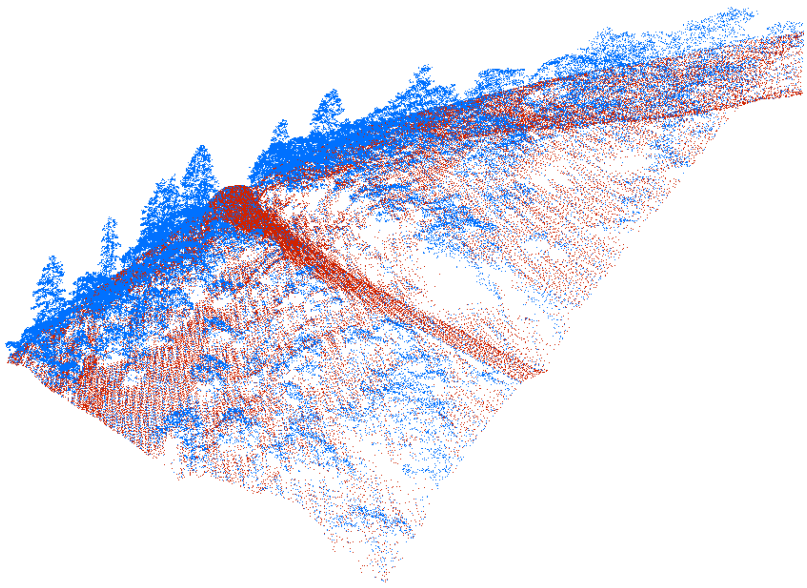


Figure 7: Lidar point-cloud visualized from Sierra Nevada including vegetation (blue) and bare-earth (red) with central projection.

Lidar provides high point sampling at very fine spatial resolutions, although the forecast from points to an interpolated grid is subject to much uncertainty. In most cases, interpolation of the point data is needed due to irregularity of the data acquisition grid (Lloyd *et al.* 2002). With respect to the object of interest whether bare-earth estimates based on DEMs or vegetation parameters (e.g., canopy height, canopy cover and building indexes) which are acquired from DSMs, accuracies of these lidar derived products depend mainly on the interpolation method, resolution and lidar point-sampling density (Bater *et al.* 2009; Priestnall *et al.* 2000; Aguilar *et al.*, 2005, Guo *et al.* 2010) (refer to *Figure 2* for interpolated vegetation points).

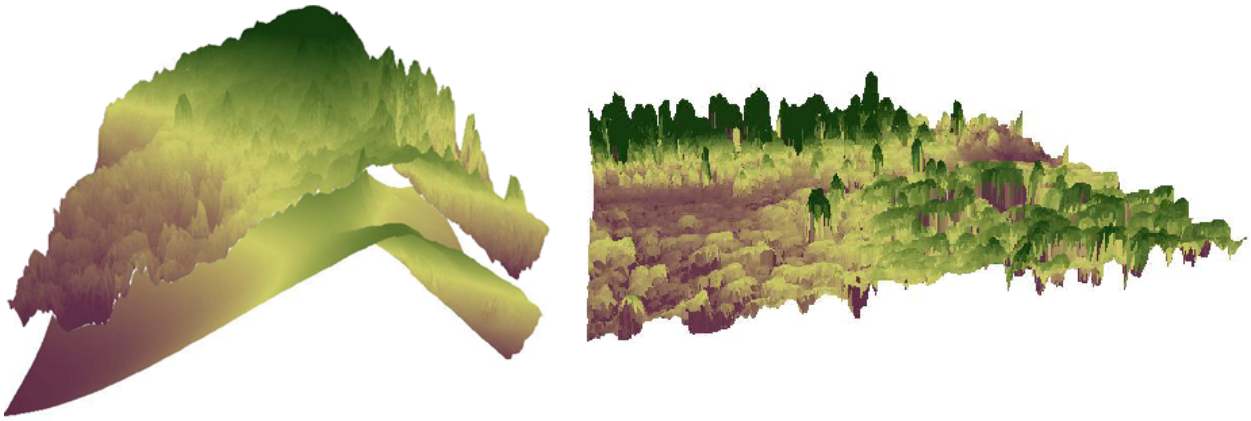


Figure 8: Top to Bottom Digital Surface Model (DSM) and Digital Elevation Model (DEM) (left). Interpolation is needed to generate these products. The Canopy Height Model (right) is acquired from subtraction of DEM from DSM.

With high resolution lidar data, DSMs are useful for landscape modeling, forest tree extraction and visualization applications *figure 3*. The quality of DSMs is also important for a variety of geographic information models and spatial processes. Studies in lidar including DEM and DSM generation include modeling flood inundation from rivers in urban environments (Priestnall *et al.* 2000) as well as studies including canopy height models (CHMs) obtained by the subtraction of the DEM from DSM and have well documented forestry applications (Xiaowei *et al.* 2004). Many studies have focused on lidar generated DEMs with respect to sampling

density and interpolation methods (Aguilar *et al.*, 2005, Guo *et al.* 2010), along with studies in interpolation methods comparison (Caruso *et al.* 1998).

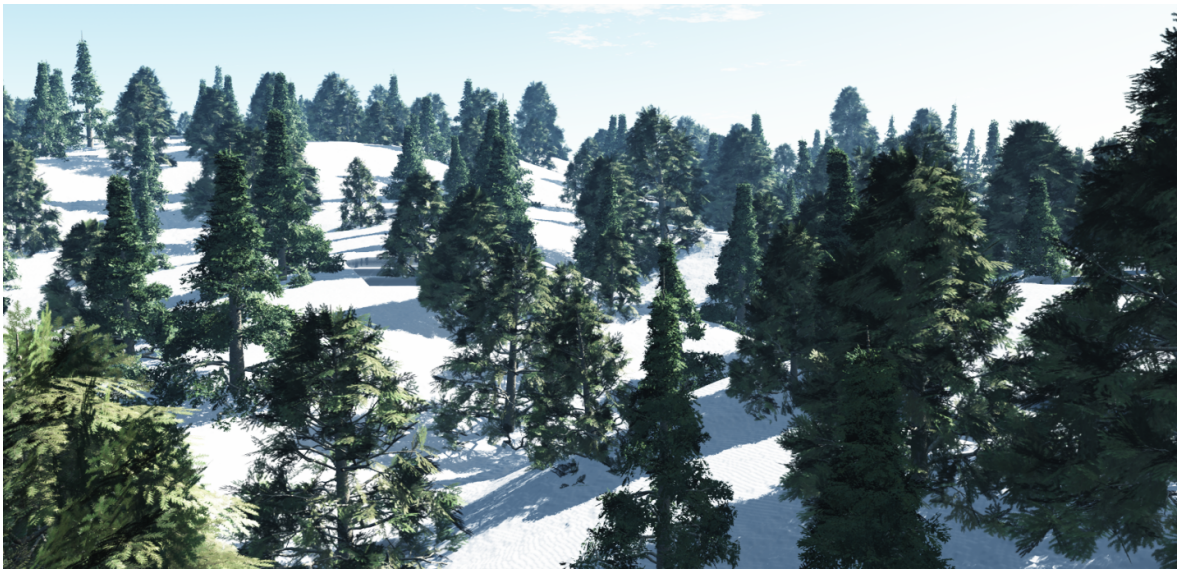


Figure 9: Lidar visualization image courtesy of Jacob Flanagan generated by Vue of the Sierra Nevada. Snow-depth was a parameter added into the image on top of DEM and individual tree detected CHM.

Lidar has proven to be a promising technology for estimating forest biophysical parameters, but due to the high-cost of flights lidar-data is not available for many geographic areas. Since lidar is also computationally expensive to process with large disk space size, lidar is difficult to process at high resolutions. One of the main limitations of lidar studies that are examined in this work is the excessively long data processing times, especially for DSM generation due to a large number of points which must be interpolated (> 1 billion for reasonably sized study areas). The majority of lidar data in this study were processed used Intel Quad Core™ technologies along with simulated parallel processing. One way to process the large amounts of lidar data is to execute separate scripts across separate central processing units (CPUs). Server and PC Random Access Memory (RAM) used in this lidar study were increased to 24 GB. In addition, multi-terabyte hard-drives (both internal and external totaling 16 TB) were installed to manage the large data sets and post-processed gridded surfaces. A USB 3.0

PCI card was also installed to increase the efficiency of data transfer between hard-drives on a Dell PowerEdge R900 windows server. The facilitation of lidar data sharing is another issue spatial ecologists and GIS specialist's face on a regular basis and will be addressed in ***Chapter 5***.

The goal of this research is to provide a framework for generating and analyzing lidar-derived product such as Digital Elevation Models (DEMs), Digital Surface Models (DSMs) and lidar-derived biomass estimates for a study area in the Sierra Nevada based in support of the Sierra Nevada Adaptive Management Project (SNAMP). SNAMP is a joint effort by the University of California, state and federal agencies, and public stakeholders to study management of forest lands in the Sierra Nevada. The goal of SNAMP investigators is to develop, implement and test adaptive management processes by testing the efficacy of Strategically Placed Landscape Treatments (SPLATs) across four response variables, including: (1) public participation, (2) wildlife (focusing on the Pacific Fisher and the California Spotted owl), (3) water, along with (4) fire and forest health. This thesis also provides an open-source framework for storing and sharing spatial data using an online web-content management system or digital library (dl). The web-based software in this study is used to store and share data amongst a variety of teams and persons including the public, the Sierra Nevada Adaptive Management Project (SNAMP) and the National Critical Zone Observatories.

Results described in this thesis include USGS and lidar-derived DEM error across multiple resolutions, DSMs generated across a variety of platforms including point-density reduction, interpolation methods and resolutions, and a comparison of forest biomass estimations using individual tree extraction from lidar and regression approaches using ground-truthed data. Individual chapters in this thesis will be submitted for peer review and possible publication to the *International Journal of Remote Sensing* and other high-impact journals.

1.1 – Study Area, Lidar Data and Ground-Truthing

The study area of interest is located northeast of Auburn, California in the Tahoe National Forest and encompasses 107 km^2. The average elevation of this study area is 1559 m with standard deviation of 293 m. The National Center of Airborne Laser Mapping (NCALM) at the University of Florida was contracted to survey the area using an Optech GEMINI Airborne Laser Terrain Mapper (ALTM) mounted on a twin-engine Cessna Skymaster. This survey was performed in five flights: one on September 18, 2008 (calendar day 262), two on September 19, (263) one on September 21 (265), and a final flight on September 22, 2008 (266). This site was chosen because: 1) Active United States Forest Service (USFS) management plans are currently in place there, 2) The location met a range of scientific criteria (including providing habitat for wildlife species and the potential for recruiting large tree structure), and 3) It is representative of Sierran landscapes.

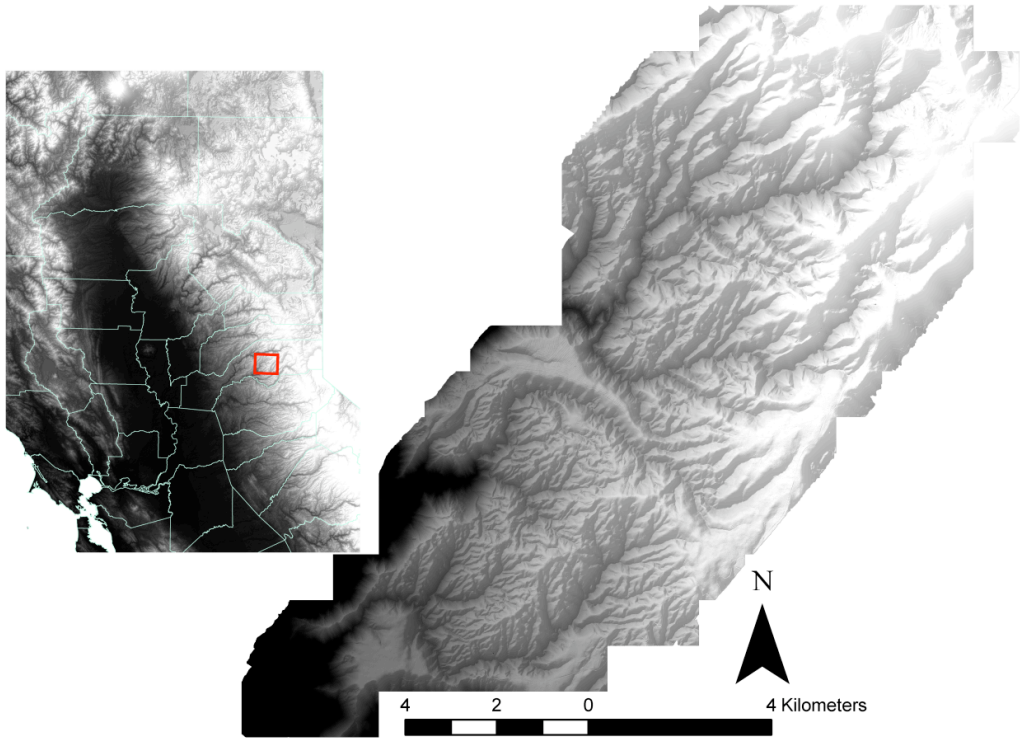


Figure 10: Northern Study area Last Chance in Northern Sierras, east of Lake Tahoe in the Tahoe National Forest.

Multiple parameters were acquired from the ground-truthed inventory for individual trees for the particular plots including; species, diameter-at-breast height (dbh), tree-height, snag, and other parameters used in other facets of the SNAMP project. Tree dbh was measured with a standard measuring tape, tree height was measured using Vertex Ultrasonic Hypsometer, and Global Positioning (GPS) of each individual tree was collected using a Trimble GeoXH. The majority of the species found in these this site vary from softwood species, hardwood species and woody shrub species along with snags, dead, stumps, and burned trees. Each plot (or area of ground-truthing), has a plot center that measures a radius of 12.62 m. In accordance with ground-truthing standards set by SNAMP, each and every tree within a 12.62 m radius from a randomly defined plot center was measured excluding trees on or under 2 m height.

The study area ground-truthed inventory included 115 plots and 2186 trees of eight different species: (*Abies-magnifica* Red Fir, *Abies-concolor* White Fir, *Calocedrus-decurrens* Incense Cedar, *Pinus-contorta* Lodgepole Pine, *Pinus-lambertiana* Sugar Pine, *Pinus-ponderosa* Ponderosa Pine, *Psuedotsuga-menziesii* Douglas Fir, and *Quercus kelloggii* Black Oak). The ground-truthed data were collected by UC Berkeley's SNAMP Spatial Team. The distribution of plots collected from the ground-truth data were stratified based on vegetation type: conifer, mixed and deciduous trees with sparse, medium and dense canopy density were ground-truthed. A full report on ground-truthed inventory data can be found in the ***Appendix*** section of this thesis.

CHAPTER 2- Lidar and USGS derived DEM Comparison

2.1 – Objective

Many hydrologic, vegetation science, and urban planning applications use digital elevation models (DEMs) to obtain absolute surface elevation and terrain form (e.g., slope, aspect) information (Jensen *et al.* 2000). DEMs may be produced using *in situ* measurements, photogrammetrically derived measurements from stereo-correlation and aerial surveys, lidar laser measurements, and interferometric synthetic aperture radar IFSAR active microwave measurements. Some studies have suggested that the accuracy of DEMs vary depending on land cover and slope. This suggestion is based on the assumption that any cover that has a substantial canopy will inhibit a visual modification of the DEM or an automatic terrain extraction algorithm. It is not known what the accuracies are for DEMs derived over certain land cover classes or whether the errors are significantly different between land cover categories (Bolstad *et al.* 1994; Hogdson *et al.* 2003; Smith *et al.* 2004; Hodgson *et al.* 2004; Hodgson *et al* 2005).

This study aims to quantify the effects of slope, canopy cover, canopy height and land-cover across multiple resolution DEMs for California and the Sierra Nevada. DEMs used in this study are from the United States Geological Survey (USGS), and lidar-derived DEMs from the study site as described in *section 1.1*. USGS derived DEM resolutions include 30 m and 10 m per pixel, while lidar derived DEM resolutions include 1 m per pixel. The USGS products were compared by means of aggregation, 10 m and 30 m, while the USGS 10 m and lidar-derived 1 m products were compared separately. The LandFire dataset was used to acquire canopy cover, canopy-height and landcover information for the 10 m and 30 m comparison. Results indicate a significant difference between USGS derived DEMs, and lidar derived 1 m products. Landcover type, and slope play a major role in DEM accuracy for all generated DEMs. Error between

canopy cover and height values did not seem to affect DEM generation significantly. The following research questions were addressed in this study:

i. Is the USGS 30 m DEM derived from the USGS 10 m product?

ii. How does slope and canopy cover correlate with DEM error between resolutions?

iii. How does canopy cover and canopy height affect USGS DEM generation?

iv. Does landcover play a major role in DEM generation?

v. Where are the over/under-detections with respect to slope and canopy cover?

vi. How can we use categorical data to evaluate the significant differences and errors?

One study in particular has evaluated the accuracy of USGS DEMs as well as lidar generated DEMs (Hodgson *et al.* 2003). This study evaluates the accuracy of USGS DEMs, lidar and ifsar over a controlled watershed with leaf-on conditions. The main contribution of this particular study is the collection of ground referenced information in comparison to these products and rigorous error assessment. Although, no research to date has evaluated the accuracy of USGS DEMs for the entire state of California and the Sierra Nevada for USGS 30 m, USGS 10 m and lidar-generated DEMs as well as evaluate the topographic error associated with canopy cover, canopy height, and landcover type across all three remote sensing DEMs. The unique contribution of this paper is the identification of error in DEM products which are highly used by many people for a variety of purposes including research and hydrological modeling specifically in the Sierra Nevada.

2.2 – Methods

The USGS has been a major producer for DEMs for the United States over the past 30 years. USGS has four primary methods of deriving DEMs that are categorized into 3 different levels of quality, Levels 1, 2 and 3. The four methods for producing these DEMs are 1) manual profiling 2) automatic correlation, 3) contour-to-grid interpolation, 4) integrated contour to grid interpolation. Each method has its advantages and disadvantages that can result in unique problems and/or artifacts in elevation products. For a complete description on USGS derived DEMs, refer to Hodgson*, et al. 2003*.

Two separate analyses of DEMs were compared for two different study area extents. The first analysis was the comparison between USGS 30 m and USGS 10 m for the entire state of California. The USGS 10 m DEM product is assumed to be a better means of recorded elevation since it has much higher resolution and results from a better method of production. To compare which topographic factors affected the generation of the USGS 30 m, the LandFire vegetation layers including landcover were used for the entire state of California. LandFire is a shared project between the U.S Department of Agriculture Forest Service and the U.S Department of the Interior wildland fire management programs and is sponsored by the Wildland Fire Leadership Council (LANDFIRE Data Products, 2011). LandFire Zones included *Zone 3, 4, 5, 6 and 13* which cover the entire spatial extent of California at 30-meter spatial grid resolution and are developed using geo-referenced field plot data, satellite imagery and simulation models. *Figure 10* displays LandFire datasets for the entire state of California including canopy cover and canopy height layers. The USGS 10 m and lidar-derived 1 m DEMs for the study area described in *section 1.1* were compared for the Sierra Nevada. Since lidar data provides high point sampling at very high spatial resolution, this remote sensing dataset is assumed to be more

accurate for estimating elevation than the USGS 10 m DEM. To acquire topographic factors in this separate analysis, the canopy height model (CHM) was acquired from interpolated lidar datasets of the subtraction between DSM and DEM. Canopy cover was calculated for vegetation greater than 2 m, if the canopy height is greater than or equal to 2 m for the particular 1 m pixel, canopy cover yield is 100%, and otherwise canopy cover is 0% (bare-earth). Lidar derived elevation, canopy height, and canopy cover was aggregated from 1 m to 10 m spatial resolution. Slope products for both the 30 m and 10 m comparisons were derived from USGS 10 m and lidar-derived 1 m DEMs respectively. ArcGIS 10 and python 2.6 were used to subtract, aggregate and project the multiple DEMs at common resolutions and also create slope, canopy cover and canopy height products. Refer to *figure 5* for the complete flow of processing for the comparison of USGS 30 m, USGS 10 m and lidar derived DEMs. USGS 10 m and 30 m products were also compared using this method in ArcGIS and python using the ArcPy library.

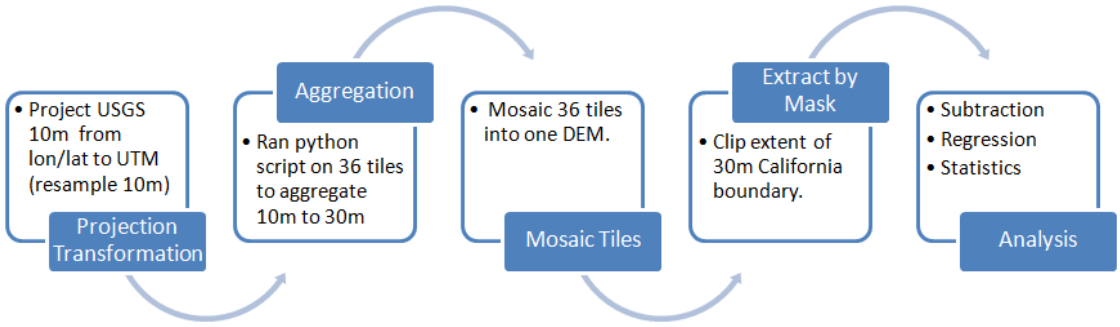


Figure 11: Processing steps for aggregation, projection and analysis of USGS 10 and 30m derived DEMs.

Once datasets of all geospatial information were created in ArcGIS including the absolute subtraction layer between elevation products, a final product (refer to *figure 6*) of combined layers was needed in order to evaluate which topographic factor(s) have the most influence on remotely sensed elevation datasets. To combine datasets into one database, the ArcPy library

SearchCursor feature was used. A cursor is a data access object that is used to iterate over the sets of rows in a table within a shapefile. In this case read-only access was used to execute a SearchCursor function over all geospatial datasets and combine each pixel in each geospatial dataset for the comparison of all remotely sensed elevation products. Since there were two comparisons, two databases were created: 1) USGS 30 m and USGS 10 m DEM comparison 2) USGS 10 m and lidar-derived DEM comparison. The Absolute Error Layer contains the absolute difference between both comparisons on pixel-by-pixel subtraction. Over and under-detections from elevation surfaces Z_{real} - $Z_{observed}$ is observed where Z_{real} is the higher resolution elevation surface (the more accurate method of generating an elevation surface) used in the comparisons and $Z_{observed}$ is the DEM under evaluation (the lower resolution elevation surface, or the method under evaluation). In case 1: USGS 30 m DEM = $Z_{observed}$, USGS 10 m DEM = Z_{real} and in case 2: USGS 10m = $Z_{observed}$, lidar-derived DEM = Z_{real}. Since adequate landcover products typically do not exist for high-spatial resolutions, landcover was not a topographic factor used in the USGS 10 m and lidar-derived DEM comparison.

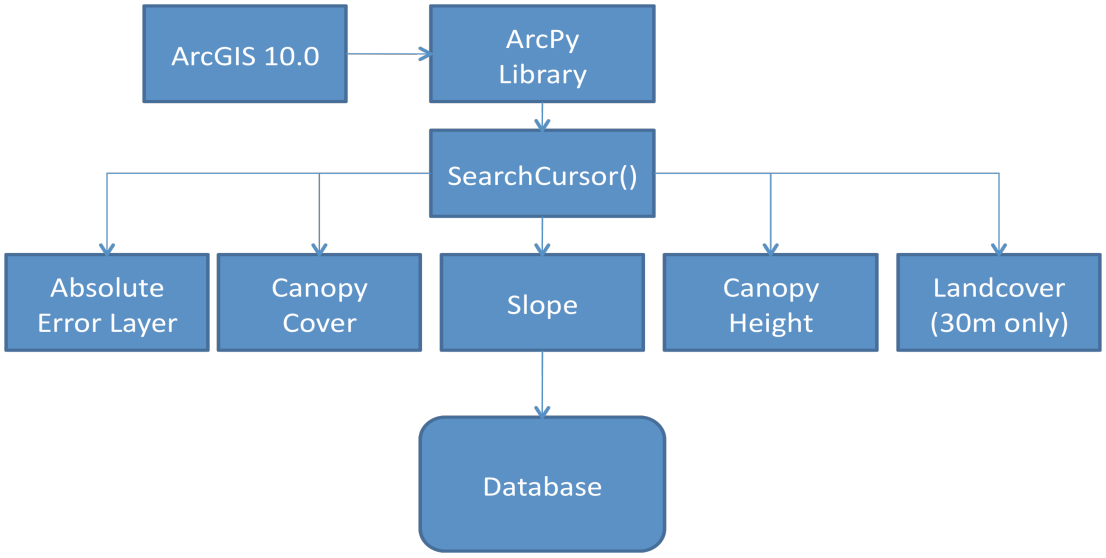


Figure 12: SearchCursor function in ArcPy Library from ArcGIS 10.0 used to combine geospatial information per pixel into database.

2.3- Results

The first comparison was made between USGS 10 m and 30 m DEMs. From the pixel-based subtraction (after the USGS 10 m elevation surface has been aggregated to 30 m), it is evident that the two DEMs are significantly different, thus proving the fact that the USGS 30 m product was not derived using the same methodology as the 10 m product. Since there was a greater over-estimate in elevation in the USGS 30 m product than that of the USGS 10 m product, this might suggest canopy properties influencing creation of DEMs based on results described in Hodgson *et al.* 2003. However, after running an analysis of variance (ANOVA) with error value (subtraction value in meters) as the *dependent variable* and the following factors as *independent variables*: Slope (degrees), Vegetation Type (LandFire vegetation key), Canopy Cover (0-100%), Canopy Height (meters), and Over/Under Detection/Zero; results suggest that slope had the greatest influence in DEM error (refer to *Figure 7*).

Table 1: ANOVA of 10 m and 30 m USGS comparison DEM including vegetation layers from LandFire datasets

Source	Type III Sum of Squares	d.f	Mean Square	F	Sig.
Slope	231643	57	4063	783	0.01
Landcover	3195	28	114	22	0.01
Canopy Height	867	5	173	33	0.01
Canopy Cover	191	9	21	4	0.01
Detection	52	2	26	5	0.01

Based on the *f-score* values acquired from the ANOVA analysis, landcover and canopy-height are also contributing factors to USGS 30 m in elevation accuracies. Canopy Cover and over/under detection in DEMs might not be a significant contributing factor to USGS derived DEMs but shouldn't be completely removed from overall DEM accuracies because this factor

might suggest dependencies on accuracies with respect to which methodology or remote sensing technique is used to derive an elevation surface. There is also a spatial-autocorrelation factor associated with elevation accuracies which will be later explained in the *discussion* section of this chapter.

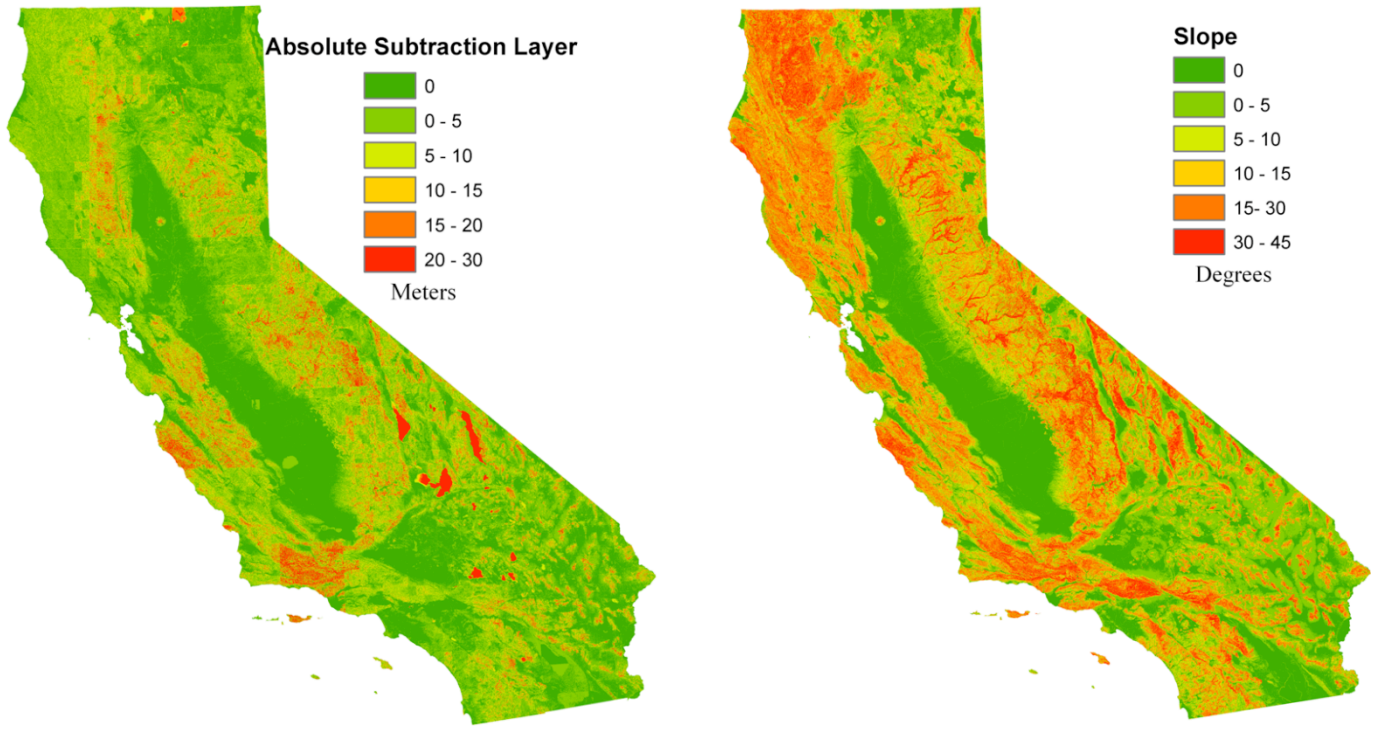


Figure 7: USGS 10 m & 30 m absolute error (left), Slope of USGS 10 m of California (right).

USGS 10 m and 30 m derived elevation errors depend mainly on slope, and canopy-height. Slope and canopy-height in error analysis were categorically binned using the following slope-values in degrees as shown in *Figure 9*. Slope yielded an exponential increase in elevation uncertainty from the USGS 30 m product. Also, as canopy-height increases, it shown that error increases linearly. On a pixel-by-pixel average, the average uncertainty USGS 30 m contains in elevation accuracy is 7 m. For higher sloped and taller canopy regions the average elevation error can be as great as 30 m. This result should encourage users of the USGS 30 m product to

use higher-resolution DEMs due to the significant uncertainties in sloped or high elevated regions. Elevation uncertainties from gridded surfaces in low-sloped areas are relatively low.

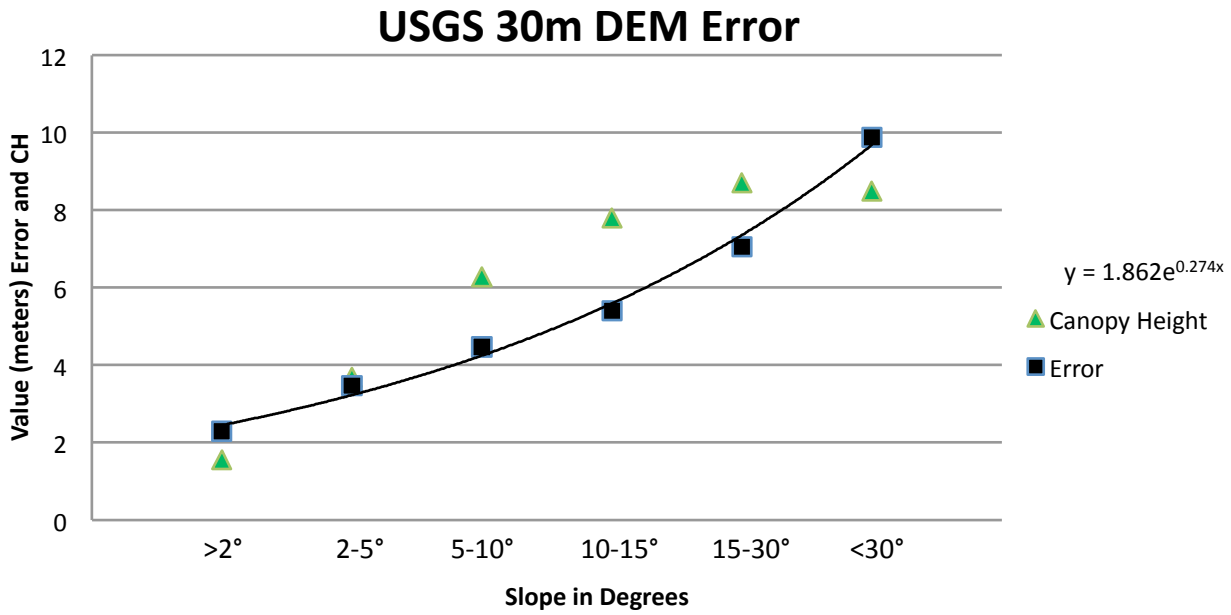


Figure 8: DEM Error increase with respect to slope and Canopy Height of USGS 30 and 10m DEM of California.

From total statistics tallied from the USGS 10 m and 30 m database the following landcover types (acquired from LandFire) should be noted on extreme outlying accuracies in USGS 30 m DEM:

i. Greatest under-detection in open-water, inter-mountain basins, deserts and lakes.

ii. Greatest over-detection in SubApline, Oak and Mesic Conifer Forests.

iii. No significant difference in Central Valley of California (i.e low-slope area).

Figure 9 displays subtracted values (error) from USGS 30 m which was compared to aggregated USGS 10 m, including greatest under and over-detections with respect to landcover. Also note over-detections in the Sierra Nevada and coastal mountain ranges. *Figure 10* also displays canopy characteristics acquired from the LandFire datasets.

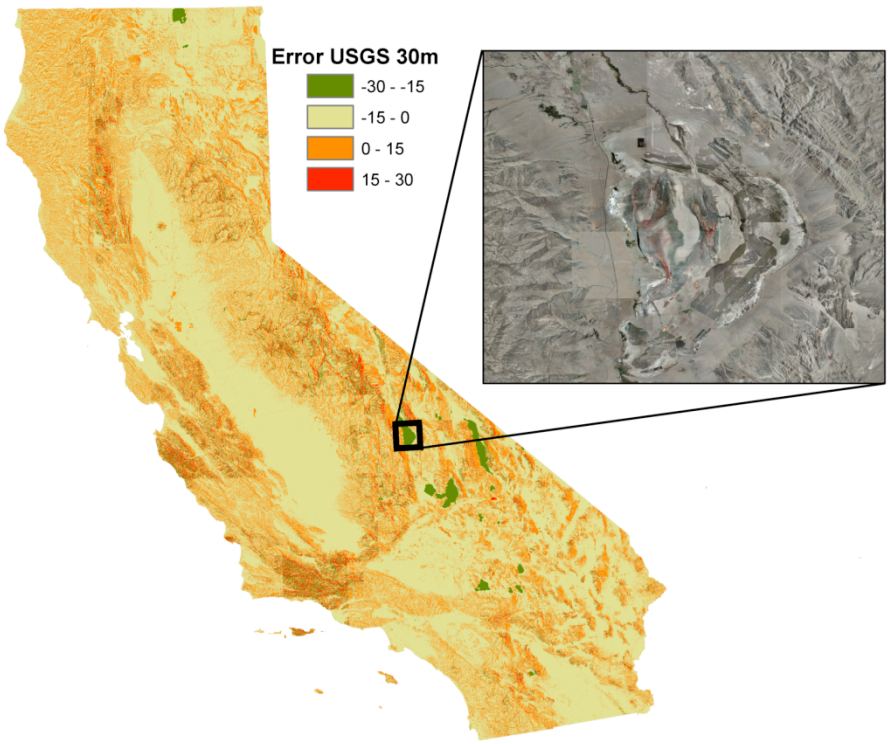


Figure 9: Subtraction Layer for USGS 30 m and USGS 10m shown. Greatest under-detection in open-water, inter-mountain basins, deserts and lakes. Image acquired from *1m* NAIP imagery.

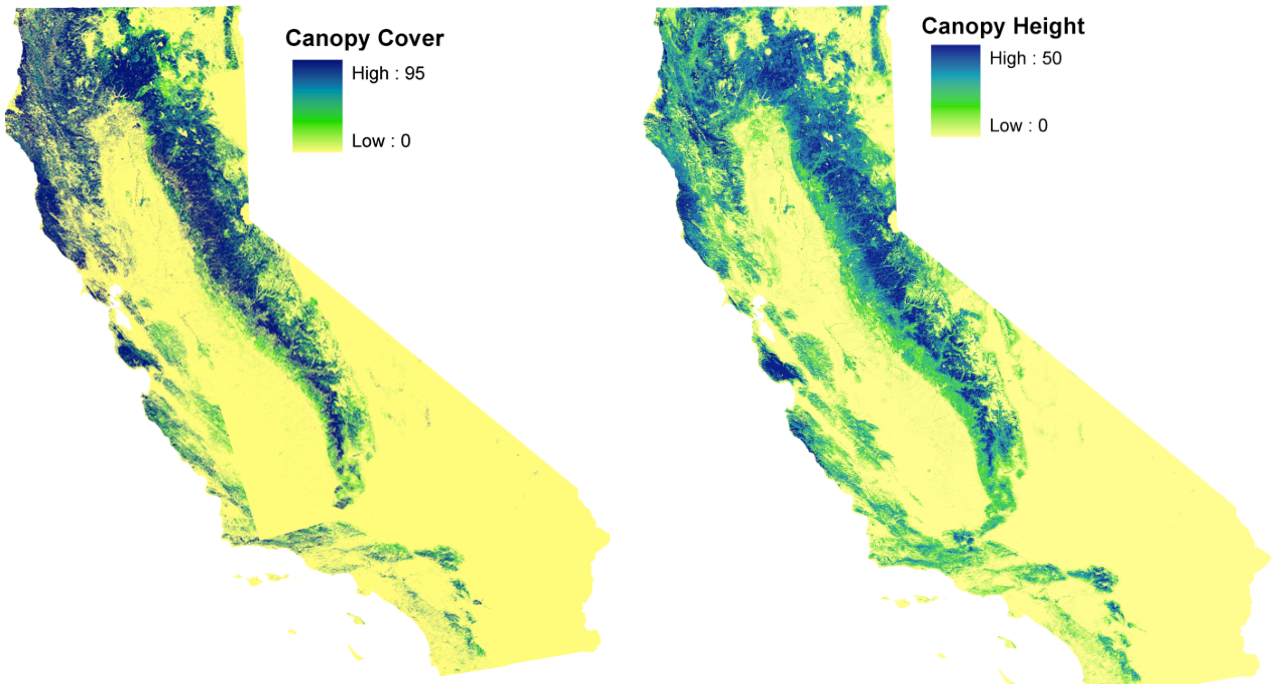


Figure 10: Canopy Cover percentage (left) and Canopy Height (right) of mosaic of California based on LandFire vegetation layers.

The comparison between USGS 10 m and lidar-derived 1 m DEM for the Sierra Nevada (study area described in *section 1.1*) yields an over-detection in USGS 10 m elevation based on over more than 1 million lidar ground-points. This might suggest that only lidar can penetrate through canopy and provide adequate DEM accuracies. The next-step in this analysis was to provide a quantitative means of accessing the accuracy of the 10 m product based on categorizing error into six different parameters. The first assumption based on a variety of studies was that slope influenced DEM generation the greatest (Su *et al.* 2006; Hodgsen *et al.* 2003; Hodgsen *et al.* 2005). *Figure 11* describes the increase in error and slope based on the subtraction between USGS 10 m and lidar-derived 1 m product. As discussed, canopy-height was acquired from a subtraction of interpolated elevations from classified ground-points and vegetation points. As shown, it is evident that great sloped areas can provide elevation uncertainties of up to 50 m *Figure 12*.

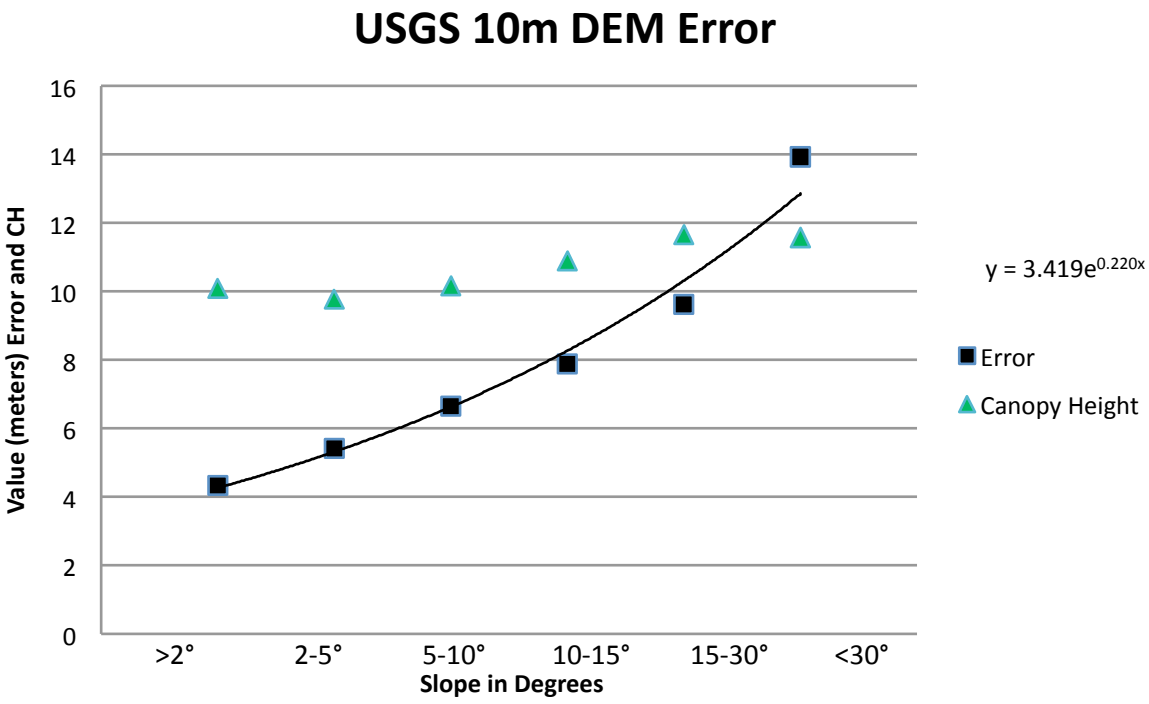


Figure 11: Exponential increase in absolute error as slope increase. Canopy Height at each categorized slope are also shown.

This analysis was also categorized based on average values from derived spatial products (*i.e.* error values, canopy-height and canopy cover with respect to slope categories were averaged). Although a separate evaluation of canopy cover was taken into consideration in comparing USGS 10 m and lidar-derived 1 m DEM, it did not provide the same trend as of error increase with slope increase even though categories were binned based on a normal distribution of canopy cover within our study area. Although based on a pixel-by-pixel analysis this proves that better elevation accuracies exist in the USGS 10 m product than USGS 30m product. Also, since canopy cover in our study area was relatively high (refer to *Figure 13*) it was difficult to quantify erroneous trends with this parameter. As previously stated, landcover was not included in this study due to insufficient coverage of vegetation products for this study area.

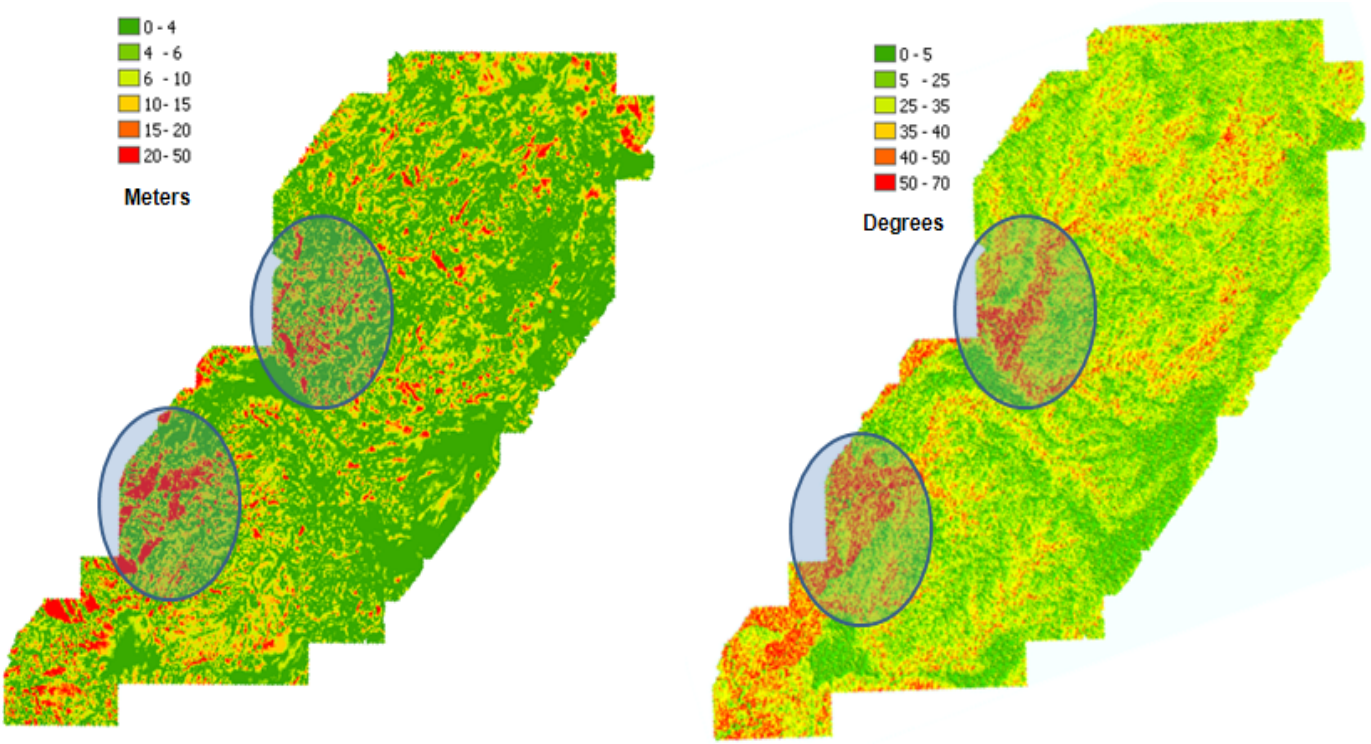


Figure 12: Absolute error in USGS 10m in comparison to lidar DEM 1m for the Sierra Nevada. Error value in meters (left) pictured next to slope in degrees (right) of study area. Highlighted areas show some regions of spatial-auto correlation.

2.4 – Discussion

As DEMs play a major role in spatial processes and modeling, it is necessary to acquire a means of error with respect to elevation and terrain attributes (Thomspon *et al.* 2001). It is suggested that when the complexity of the terrain increases with respect to topographic variable parameters that uncertainty in elevation increases (Smith *et al. 2004*). Although it is noted that it is difficult to combine all of these factors of topographic variability together into one study due to strong colinearlity between them (Guo. *et al.* 2010). Since there a variety of methods to represent topographic variability, the inclusion of fractal dimension, semivariogram, coefficient of variation and elevation variation were not included in this study. It should also be also noted that for the analysis of USGS 30 m and USGS 10 m for the entire state of California, that spatial-auto correlation exist for the Sierra Nevada and mountain regions due to: highly sloped, dense canopy (refer to *Figure 13*), and tall vegetated regions as well as distinctive landcover features. As opposed to the mountainous regions in California, the Central Valley is a low sloped, sparse canopy and low vegetated region with homogenous landcover characteristics. The Central Valley has low uncertainty in elevation accuracies in both USGS 30 m and 10 m DEM products. The inclusion of snow-on conditions or water levels may lead to over-estimate of elevation especially in mountainous regions, this suggest that time-of-year in acquired elevation surfaces must be recorded during a time of year where snow levels are low and mountainous regions present leaf-off conditions for bare-earth penetrability from remote sensors.

The relationship in reported elevation is strongly related to slope, canopy-height and landcover properties for Case 1: USGS 30 m and 10 m comparison. Although, in Case 2: USGS 10 m and lidar-derived 1 m comparison, there was not a clear trend in canopy parameters affecting erroneous elevations. The suggestion of canopy inhibiting lidar sensors from

penetrating ground is a measurement that was very difficult to quantify in this study due to high density vegetation and canopy cover within study area without the inclusion of ground-recorded elevation accuracies. The suggestion of canopy inhibiting elevation accuracies also depends on methodologies of deriving elevation surfaces as described in Hodgsen *et al.* 2003.

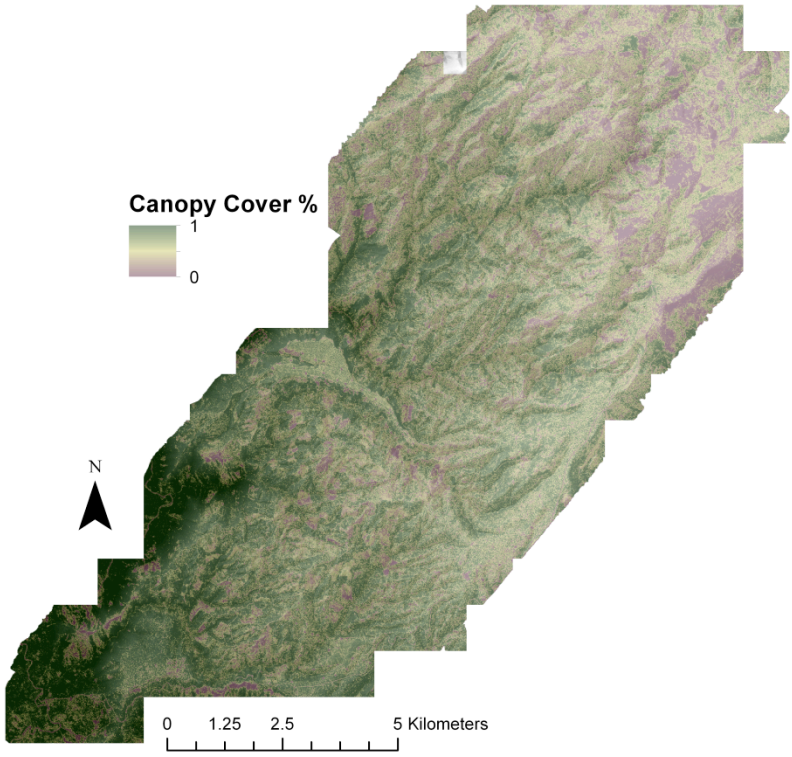


Figure 13: Canopy Cover percentage for study area.

LandFire vegetation products that were used in this study as a means of vegetation layers are mapped using predictive landscape models based on extensive field-referenced data, satellite imagery, biophysical gradient layers, and classification and regression trees. These data are useful for determining existing vegetation conditions, for change detection, and for natural resource management analysis. Users of LandFire datasets should also consider the amount of error in all of their products as a comparison of this data to lidar generated products displayed

statistically significant amounts of error. A comparison of the LandFire product to lidar-derived products for the study site was beyond the scope of this work, but warrants further investigation.

The main observation in acquiring accurate elevation and using accurate elevation models suggest slope, and vegetation type (landcover), as described in previous studies (Smith *et al.* 2004; Hodgsen *et al.* 2003; Hodgsen *et al.* 2004; Hodgsen *et al.* 2005). It should be noted mainly that high sloped regions can be erroneous in both USGS and lidar-derived DEMs, although great errors of under and over-detection of differences between DEMs should be taken into consideration when acquiring elevation from large spatial areas where lidar data becomes unrealistic to process.

Since the ANOVA analysis in this study was solely based on *f-score*, it is not completely valid to assume parameters in this study based on higher *f-score* were high contributing factors unless a *partial eta squared* analysis is able to be performed. Since the *partial eta squared* analysis was not included in this study as in Hodgsons *et al.* 2003 it is difficult to make a clear distinction on whether canopy-height or landcover variables affect DEM generation the greatest. Although a greater *f-score* generally yield a higher *partial eta squared* which deems that particular variable in the ANOVA analysis to be a contributing factor on the dependent variable (in this case the dependent variable was error with respect to the subtraction of DEMs as stated in *section 2.3*).

Also, when generating DEMs from lidar data it is important on which interpolation method is used, with other parameters suggested in publications (Christopher W. Bater, Nicholas C. Coops, 2009; G. Priestnall, J. Jaafar, A. Duncan, 2000; Aguilar *et al.*, 2005, Guo *et al.* 2010) including: topographic variability including canopy cover, slope, and coefficient of variation. Although, choosing the appropriate interpolator can have its own advantages and disadvantages

with respect to terrain complexity and topographic variability. In this study, the Universal Kriging method was used to interpolate the raw lidar data at a spatial resolution of 1m, although it is suggested the high-resolutions (*such as 0.5m*) predict elevation accuracy with less uncertainty. Further research is needed in order to explore what best parameters of topographic variability have the greatest influence of elevation accuracy.

One limitation in processing large amounts of spatial data in ArcGIS is the 2 GB limit of shapefiles. This proved to be a limitation is using such large datasets such as the lidar or USGS/LandFire data for the entire state of California. One way around this is to use Data Cursors in ArcPy and extract rows from each feature class into a text-file as described in *section 2.2*. Python can then be used to combine all parameters in the study into one large text file and then imported into statistical software to complete the analysis. The combination of text file size including USGS and lidar-derived parameters in this study did not exceed 50 GB. Both DSM and DEM were interpolated using Universal Kriging at a spatial resolution of 1 m for the study area. SPSS was used to generate descriptive statistics on the datasets. Intel Quad Core CPUs were used to process the majority of the data to decrease computation time due to the large size of files, especially from lidar datasets.

CHAPTER 3- Deriving Digital Surface Models from Lidar

3.1 - Objective

Point cloud density, spatial resolutions and interpolation methods play a major role in lidar derived Digital Elevation Models (DEMs) and Digital Surface Models (DSMs) as stated in previous chapters. DEMs are created by interpolated classified ground-points, and DSMs are created by interpolated classified vegetation points, excluding ground objects. In this study, DSMs were generated from full density lidar points and reduced to 80%, 60%, 40%, 20% and 5% of original lidar point density (20 pts/m^2 on average). Lidar resolutions were generated at 0.5m, 1m, 5m, and 10m using interpolation methods: Triangulated Irregular Network (TIN), Inverse Distance Weight (IDW), Spline, Original Kriging (OK) and Universal Kriging (UK). The DSMs generated across these different platforms were compared using a 10-fold cross validation method and the Root Means Square Error (RMSE) from original points to the interpolated surface. Results based on 3-Way Analysis of Variance (ANOVA) suggest that interpolation method along with resolution have the greatest impact on lidar derived DSMs. Data density reduction proved to have a small significance in generating lidar derived DSMs. UK, OK and TIN proved to be the best interpolation methods at a resolution of *0.5 m*.

A very small portion of studies have focused on interpolation methods with respect to DSM generation. Two studies in particular assess kriging methods and their affects on lidar derived DSMs (Lloyd, C.D. and Atkinson, P.M., 2002) as well as using universal kriging interpolation approach in lidar error (Coveney *et al* 2010). Lidar derived error from DSMs have also been studied and linked to different interpolation methods for urban areas (Smith, S.L *et al.* 2004). Although some generated lidar DSMs include ground points (Brovelli *et al* 2004), to

eliminate complications in this thesis, DSMs in this study are interpolated vegetation points or earth's surface excluding all ground objects.

3.2 – Methods

The following interpolation methods used in this study are described below:

Triangulated irregular network (TIN) is an alternative terrain representation approach that partitions a surface into a set of contiguous, non-overlapping triangles (Polis and McKeown, 1992). Elevation is then recorded for each triangle node, while elevations between nodes can be interpolated, thus allowing the generation of a continuous surface.

Inverse distance weighted (IDW) is a simple interpolation method that estimates the value of a point by averaging the values of sample data points within its neighborhood. This interpolation method is based upon the geographic principle that objects that are closer together tend to be more alike than objects that are farther apart (Tobler, 1970); this method gives more weight to nearby points than to distant points.

The spline method estimates values using a mathematical function that minimizes overall surface curvature, resulting in a smooth surface that passes exactly through the sample points (Bojanov *et al.*, 1993). In this study we chose tension spline which needs two parameters to be defined: weight and number of points. The weight parameters defines the third derivative of the surface in minimizing the curvature expression, a higher weight allows for a smoother surface, although too high a weight produces results that lack detail. Since we ran into some issues later described with regularized spline, it was found that tension spline weight: 10.0, points: 12 yielded the best results.

Both universal kriging and original kriging were used as interpolation methods in this study. Kriging is an advanced geostatistical procedure that generates an estimated surface from a

scattered set of points with z-values. It is based on the regionalized variable theory that assumes that the spatial variation represented by z-values (in this case vegetation height), is statistically homogenous throughout the surface. To estimate the spatial variation, the semivariogram is estimated by the sample semivariogram which is computed for the input point data set. To read more about universal kriging and the specific approaches used in each interpolation method, refer to (Guo *et al.* 2010).

Every interpolation method was run at 0.5m, 1m, 5m, and 10m resolutions with 20 datasets of variability within the study sites dataset. To assess the accuracy of the interpolation methods, a 10-fold cross-validation (Kohavi *et al.*, 1995; Picard and Cook, 1997) was applied to our lidar data set. The lidar points were first randomly divided into 10 sub-samples. We retained one of the 10 sub-samples as the validation data for testing the models performance, and used the remaining nine sub-samples as training data for DSM interpolation. We repeated the process 10 times so that all sample points were used for both training and validation. Root Mean Squared Error (RMSE), a widely used global accuracy measure for evaluating the performance of DEMs (Aguilar *et al.*, 2005) was also implemented on DSMs:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Z_i^{predicted} - Z_i^{real}\right)^2}{n}} \qquad (14)$$

where $Z^{predicted}$ is the predicted surface elevation, Z^{real} is the real surface elevation from lidar ground points, and n is the total number of points. The objective of this study focuses on the interpolation errors only.

TerraSolid's Terrascan software was used to classify vegetation points after the lidar was flown. ArcGIS 9.3 tools and the arcgisscripting library in python 2.5 were used to interpolate all

point datasets into surfaces. After each interpolation method ran at the common resolutions, as well as separate data-densities, lidar points were then compared to the interpolated surfaces. The 10-fold cross-validations were combined to compute Mean Square Error (MSE) between the lidar points and the interpolated surfaces. After the MSE was computed, RMSE was computed based on the number of points for each interpolation method, density and resolution. *Figure 15* describes that data processing from raw lidar to DSM generation.

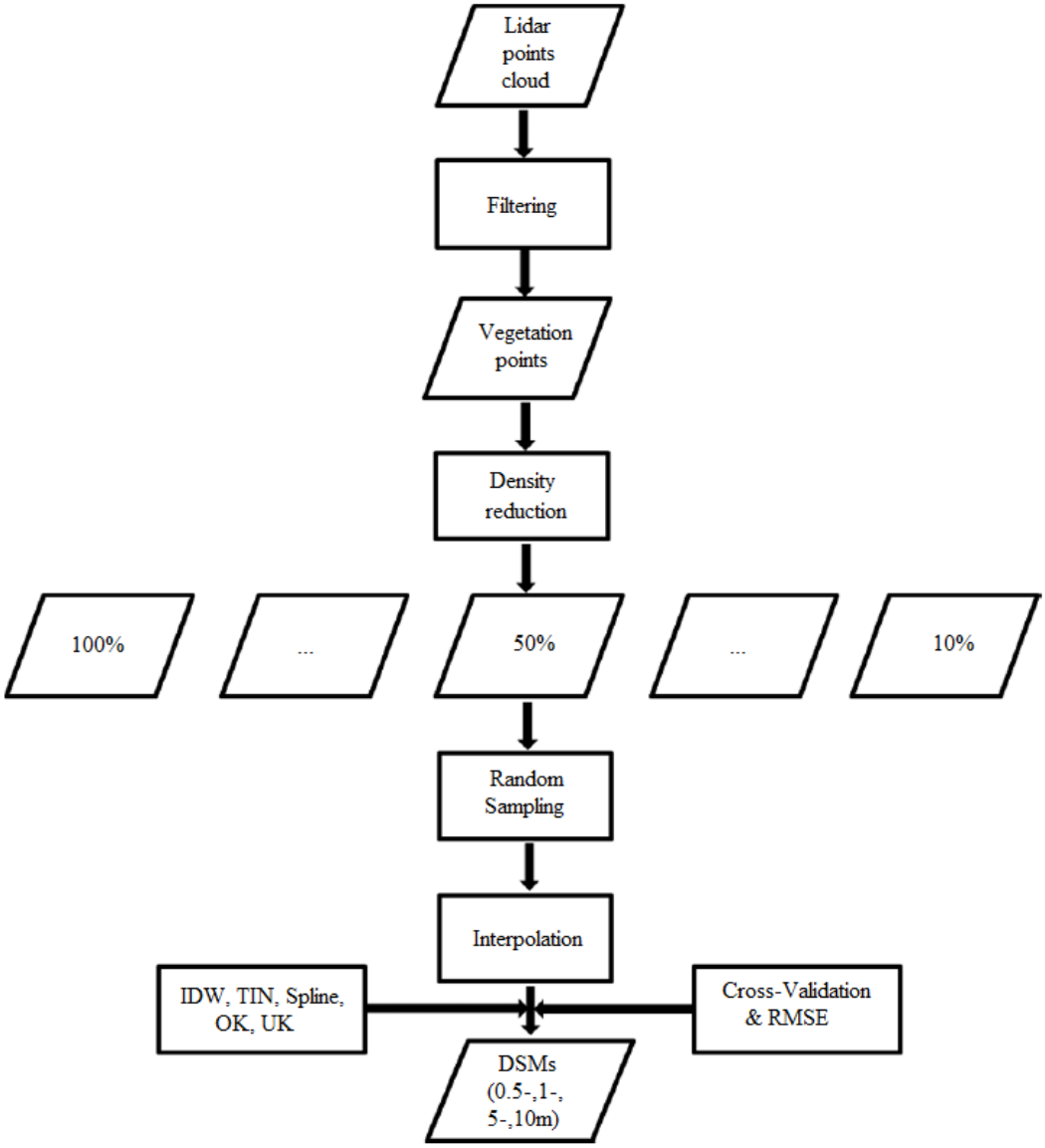


Figure 15: Flowchart of creating Digital Surface Models related to this study using lidar data density reduction, multiple interpolation methods & resolutions.

3.3 - Results

Factors in DSM generation including: resolution, density, interpolation method and interactions between these factors were compared for statistical significance at a 95% confidence level. The 3-Way ANOVA statistical test was used to determine if the means in a set of data differ when grouped by multiple factors. The comparison of RMSE values on multiple densities, multiple interpolation methods and resolutions were used to determine which factors or combinations of factors are associated with the difference. The 3-Way ANOVA is a generalization of the two-way ANOVA. In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal. Based on the higher *f-score* and RMSE trends it is safe to conclude which interpolation method and resolution has the most influence on lidar generated DSMs (refer to *Table 2*).

Table 2: 3-Way ANOVA Results for DSM data: resolutions 0.5m,1m,5m,10m at densities 5%, 20%, 40%, 60%, 80% and 100% using TIN, IDW, SPLINE, OK and UK Interpolation methods. *Significance level: 0.05

Source	Sum Sq.	d.f	Mean Sq.	F	Prob. > F
Resolution	4.08e+007	2	20240680	1621	0.01
Density	8.01e+006	4	2002886	160	0.01
Method	9.91e+007	3	33050692	2648	0.01
Resolution*Density	1.29e+007	14	924574	74	0.01
Resolution*Method	1.45e+008	11	13224487	1059	0.01
Density*Method	3.75e+007	19	1977720	158	0.01
Resolution*Density*Method	4.94e+007	59	838429	67	0.01
Error	9.43e+007	7560	12479		
Total	4.13e+008	7678			

Since the lidar data points in this study were an average of 20pts/m^2 in heavily dense canopy regions, data density reduction played a minor role in generating DSMs. Even at a data reduction of 5%, which is 1pt/m^2, it is found that DSMs can be produced without losing significant detail for particular interpolation methods. *Table 3* and *figure 16* describe the trend in errors across the multiple interpolation methods and the increase in error from fine to course spatial resolution in lidar-generated DSMs.

Table 3: Descriptive Statistics of RMSE of each interpolation method across all resolutions & densities.

Method	Mean	Median	Std. Dev	Min	Max
Universal Kriging	5.03	4.95	1.43	2.13	8.71
Original Kriging	5.78	5.68	2.19	2.03	12.45
TIN	6.05	5.78	2.4	1.07	13.26
Spline	6.85	6.33	3.47	1.29	16.31
Inverse Distance Weight	10.89	10.27	3.25	4.80	19.42

UK and OK interpolation methods yielded the lowest RMSE values through all resolutions and densities based on mean values, although these methods proved to be computationally expensive; the time required to estimate the semivariogram for the kriging method is very long for producing a surface for each tile. A detailed description of the kriging method along with problems that might arise from using kriging has also been well documented for spatial applications (Cressie, N., 1988, Cressie, N., 1990, Armstrong, M., 1984). One of the main problems with using an interpolator such as kriging is its tendencies to smooth detailed information in DSMs. TIN and tension spline produced decent results; although a visual inspection of both interpolators is required in order assess the quality of each interpolator. TIN produces better quality DSMs than spline (both tension and regularized) but during the RMSE

calculation it was noted that corners or edges of the tiles were not interpolated very well. For the corner or edge values in the DSM there were no-data values, which needed to be removed during post-processing once the DSMs were created.

As *figure 16* (vi.) displays regularized spline with parameters weight: 0.1 and points: 12, produces spikes or cones for tree tops and very high RMSE values. Tension spline is a generalization of the cubic spline and is used to avoid extraneous inflection points as well as to interpolate a surface without sacrificing smoothness. After multiple trial and error runs it was found that using a relatively high weight: 10.0, with points: 12 for tension spline produced the best results without increasing smoothness to lose detail and reduce inflection points caused by regularized spline. Statistically and visually UK and OK *figure 16* (*i. and ii*) produce the best lidar-derived DSMs, although they are very computationally expensive especially for large lidar datasets.

As noted by the 3-way ANOVA analysis, resolution is a contributing factor in the uncertainty of lidar-generated DSMs *figure 17* displays the lidar generated DSMs at multiple resolutions, 0.5m, 1m, 5m, and 10m. It is shown that as resolution increases, detail from the interpolated vegetation surface decreases. Although, this study focuses on forested regions and high-point lidar data density, this result is obvious as there is great detail in vegetated surfaces generated by lidar that is lost when spatial resolution is decreased. Since the 3-way ANOVA analysis deemed lidar point-density as less of a contributing factor to the uncertainty of point-to-surface DSMs, this parameter was explored in detail in *figure 18* and displays major factors on uncertainty involved in interpolation methods; specifically those methods that depend on neighborhood operations such as IDW, and spline.

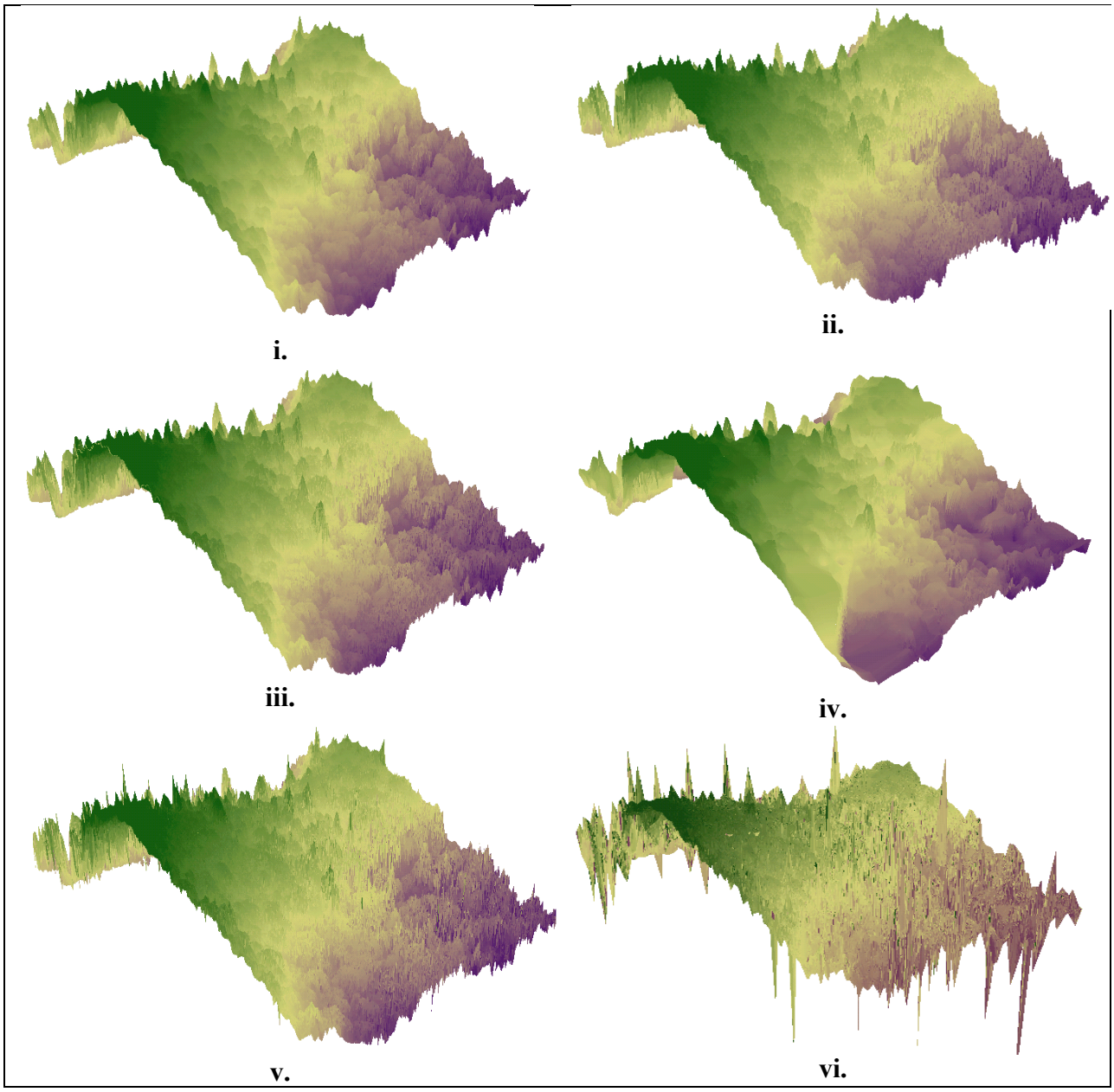


Figure 16: All interpolation methods used in this study: **i**) Universal Kriging, **ii**) Original Kriging, **iii**) Triangulated Irregular Network, **iv**) Inverse Distance Weight, **v**) Tension Spline, **vi**) Regularized Spline

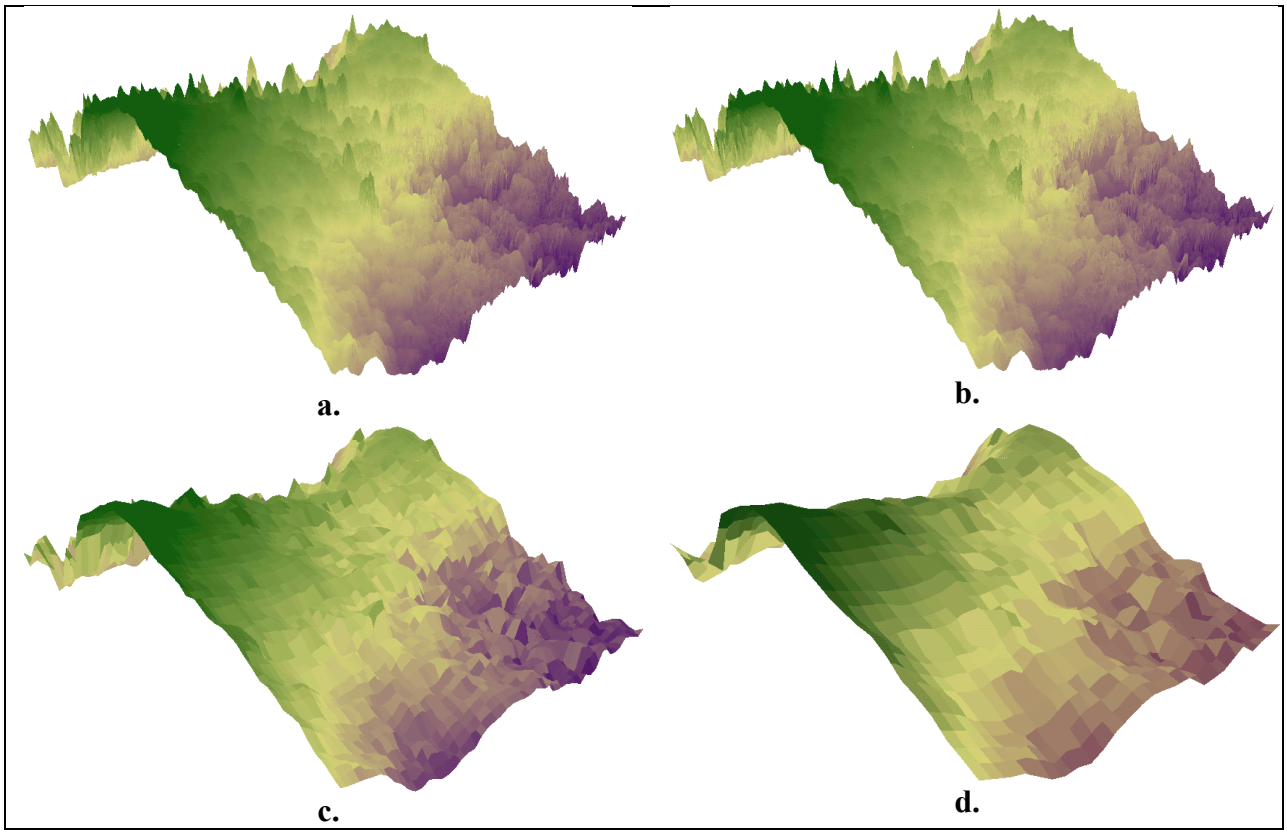
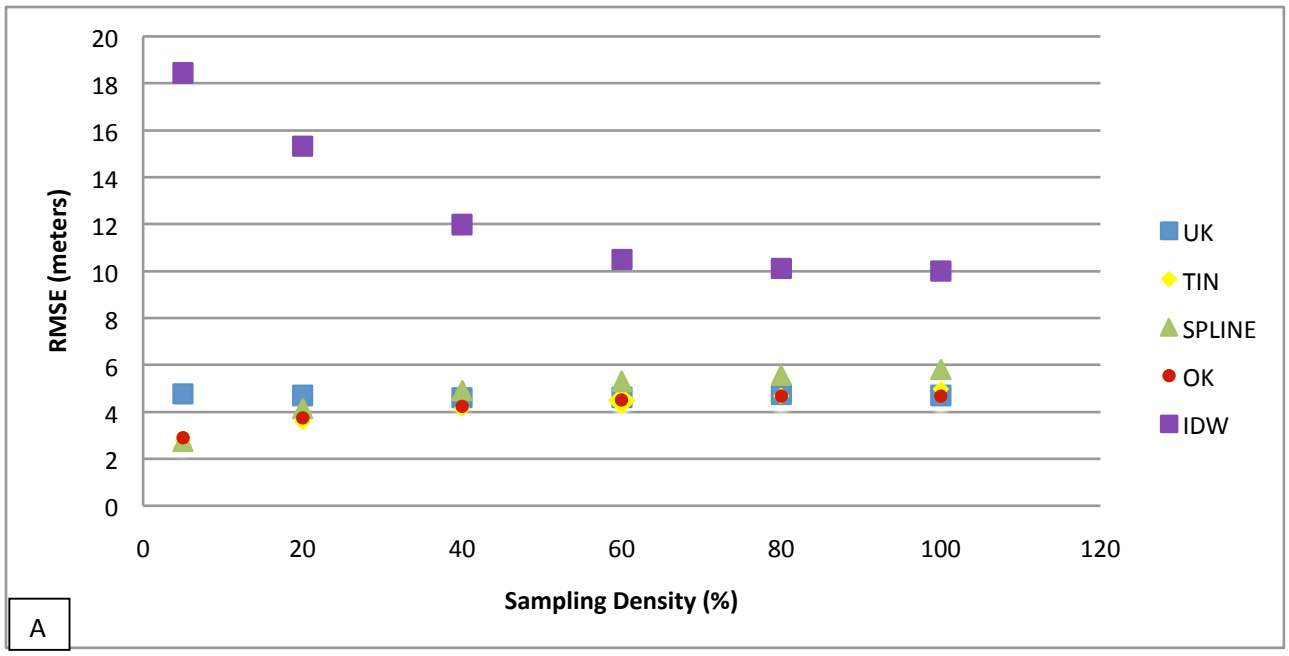
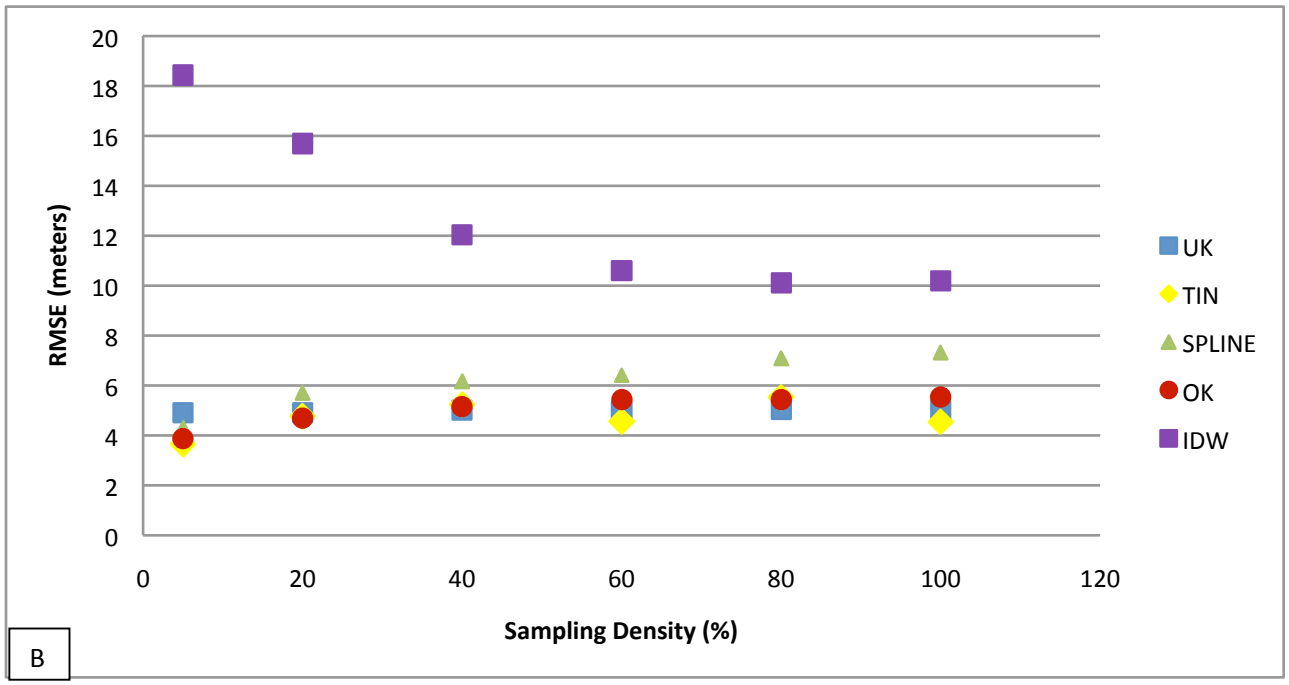


Figure 17: Multiple interpolated resolution for Universal Kriging method**: a)** 0.5m**, b)** 1m**, c)** 5m**,** and **d)** 10m

DSMs at very high sampling density (in this case 20pts/m^2 on average at full lidar data density), have greater uncertainty with interpolation methods such as spline. It should be noted that as sampling density decreases especially below 40% of original lidar points (8pts/m^2 on average), tension spline level of uncertainty is relatively low in comparison to other interpolation methods. When gaps exist between points, splines do a good job minimizing surface curvature while filling in the "holes" or "gaps" in data with lower uncertainty. Although as lidar-point data density increases, especially to the point sampling density of 60% or greater (12pts/m^2 on average), interpolation uncertainty in the IDW interpolation decreases. This decrease in uncertainty is due to the fact that IDW relies on sample data points within a neighborhood. *Figure 18* refers to all interpolation methods separated by resolution across all sampling lidar point sampling densities.


RMSE (meters)
20
18
16
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120
Sampling Density (%)
UK
TIN
SPLINE
OK
IDW
A


RMSE (meters)
20
18
16
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120
Sampling Density (%)
UK
TIN
SPLINE
OK
IDW
B

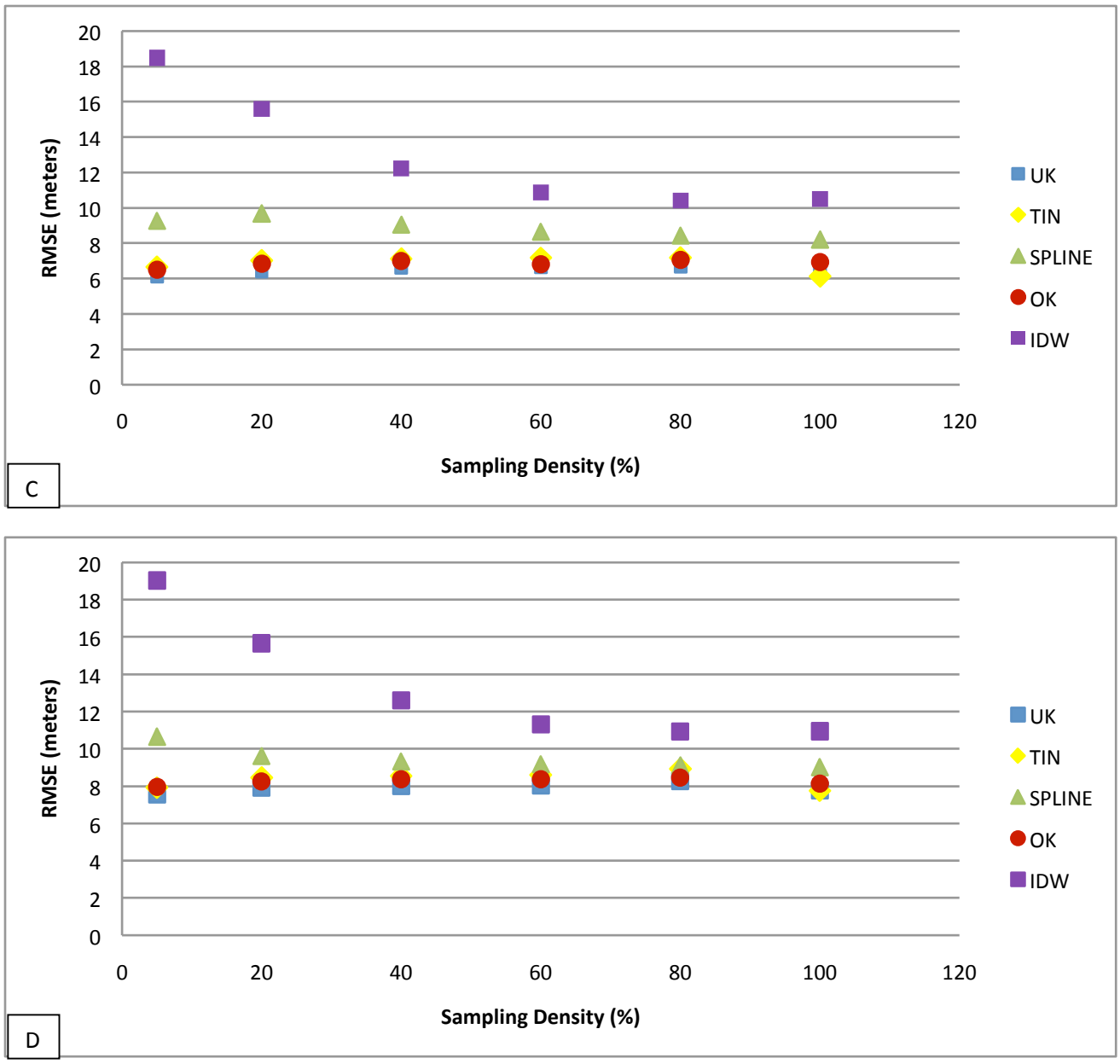


Figure 18: Relationship between RMSE and sampling density at multiple resolutions from: A) 0.5m B) 1m C) 5m D) 10m

3.4 – Discussion

The quality of lidar generated DSMs is of importance for many geospatial applications including urban and non-urban environments. Since the lidar generated DSMs in this study are a result from a densely forested area in the Sierras, it should be taken into consideration that many of these interpolation methods might vary in performance for particular areas in both urban and

non-urban environments. One of the reasons this analysis was performed was due to the many studies including published results for forestry applications using DSMs that do not describe the interpolation method or resolution used in creating point-to-grid surfaces of classified vegetation of lidar data. Since resolution increases error in lidar-derived DSMs it should be noted that a finer resolution produces better detailed surfaces, and different interpolation methods can improve or decrease both detail and elevation of surface models. Loss of information due to interpolation method or resolution can propagate error throughout products such as canopy height information and canopy cover information as well as detecting and extracting individual trees from the lidar surface which will be later in the ***Chapter 3 discussion*** section.

Since topographic variability in lidar derived surfaces especially for forested areas are somewhat homogenous with respect to landscape, additional information such as variation in canopy cover should be noted when generating lidar-derived DSMs. Canopy cover in our study area as shown in *figure 13* was is relatively dense with respect to some open areas which should also be noted in this study. Another parameter to include in this study is the coefficient of variation (CV), which is the ratio of the standard deviation of the spread of the points to the average number of points. CV of elevation (Chaplot *et al.*, 2006), is another method that describes topographic variability. There are varieties of other methods used to describe topographic variability in elevation (Guo. *et al.* 2010).

Classification algorithms, especially those of vegetation in lidar data can be another factor of providing quality DSMs. The correct filtering of lidar-data is also needed to remove "mis-hits" or mis-classifications which can occur from birds or other objects in-or-above the lidar flown area of interest. TerraSolid vegetation classification schemes should be noted in generated lidar DSMs in this study. The removal of outliers and visual inspection of any

anomalies in the lidar data should also be taken into consideration when evaluation point-to-grid interpolation.

Another issue in providing a means of deriving lidar DSMs with multiple parameters is computation time needed to interpolate raw lidar points to a surface and comparing points to interpolated surface. ArcGIS functions were used in this study such as extractByValue and other functions to compute RMSE through the dbf files and since ArcGIS is limited in the amount of processors it can access, it would be ideal to use separate GIS software to perform heavy computations and those such as lidar data using the Geospatial Data Abstraction Library (GDAL) libraries and python. The computation of this Lidar dataset ran under a Windows™ server with Intel Quad-Core 2.93 GHZ processors and 24 GB Memory using ESRI ArcGIS 9.3 and python 2.5. OK and UK generate the most accurate DSM, but their processing time is also the greatest. TIN and spline generate quick DSMs, although the use of the spline interpolation method should be visually inspected before its use especially the use of regularized spline which produced cone-shaped tree-tops as shown in *figure 16 (vi.)*. IDW had the highest cumulative RMSE value across all data densities and should be avoided due to erroneous values created in its interpolated surfaces. The comparison of point to interpolated surface was very computationally expensive using ArcGIS and python.

Choosing an appropriate interpolator has both advantages and disadvantages as previously stated. Kriging is very time-consuming, but produces the most accurate depictions of point-to-grid surfaces for forested lidar points. Splines seem to provide a relatively good interpolation technique but due to inflections and numerical instability (Guo *et al.* 2010) may not provide very accurate means of elevation from classified vegetation points. The use of an appropriate weight for splines with respect to the area of interest also needs to be taken into

consideration. TIN produces very fast interpolations with relatively good results although there is no universal approach to solving the best interpolation method due to variability of points and the error associated with point-to-grid interpolations. IDW should be disregarded in interpolating dense forested regions due to point-to-grid uncertainty.

The unique contribution and aspect of this study are high-resolution (0.5m and 1m) lidar-generated interpolated datasets for vegetation. Very few studies have focused on a variety of interpolation methods with data density reduction across multiple resolutions for surfaces generated from vegetated lidar data with respect to point data density reduction. Further in-depth analysis will be included as well as an update to the 3-way ANOVA analysis, topographic variability factors such as coefficient of variation within interpolated point-to-surface and a possibility of including other interpolators such as nearest neighbor in this study. This article will be submitted to the *International Journal of Remote Sensing* letters once completed.

CHAPTER 4 - Comparison of Biomass estimates from Lidar

4.1 - Objective

Since there has been a drastic increase in atmospheric concentrations of carbon dioxide (CO_2) and other greenhouse gases as a result from the previous century including the industrial revolution, our society is currently focusing on methods to sequester carbon to mitigate climate change (Jackson *et al.* 2005; R. Lal., 2004). Biomass (biological material from living or recently living organisms) is a renewable energy source that has the ability to produce electricity or product heat. Biomass can also generate biodegradable waste that has the ability to be burnt as fuel. Since forests are a considerable part of the global carbon cycle as they are able to sequester large amounts of CO_2, estimates of total component and above ground biomass are of importance due to the fundamental understanding of forest carbon cycles and concerns regarding climate change (Callaway *et al.* 1994). Remote sensing techniques have become more valuable in extracting parameters from the earth's surface including biomass estimates since CO_2 sequestration in high-volume biomass forests is difficult to acquire, especially for conventional and optical and radar sensors (Lefsky, *et al.* 2002). Better methods for characterizing biomass estimates from forests are sought-after to understand the overall implications of CO_2 sequestration, climate change, as well as the impact of anthropogenic disturbances including landuse and landcover changes.

There are a variety of methods used to acquire biomass characteristics from forest plots that scale at multiple ranges along with different units of measurement. The most common form for deriving forest biomass is through the use of destructive sampling and regression. In this method, trees are measured standing, and then cut and weighed. The dry mass of each trees' particular components (leaves, branches, trunks) and is then regressed by allometric equations.

Although, equations between species are sometimes interchangeable, studies show that many require trees to be similar in terms of architectures granted that allometric equation for one species to be successfully applied to the other species in the same category. Species-specific equations aren't various due to costs, labor, etc (R. M. Lucas *et al.*, 2008). Although, Jenkins *et al.. 2003* provides a framework for deriving biomass equations for a variety of tree-species in North-America.

Some studies have focused on individual tree biomass estimates using small-footprint lidar and plot-level biomass estimates (Zachary J. B, Randolph H. W, 2005; Sorin C. Popescu 2005; R. M. Lucas *et al* 2008). Although, some studies do focus on implementing the use of hyper spectral remote sensing on tree-vegetation and biomass estimates (Moses *et al.* 2007; Treuhaft *et al.*, 2003). Since multi-spectral and hyperspectral imagery were out of the scope of this study, this paper focuses solely on plot-level and individual tree biomass estimates from raw lidar, and ground-truth measurements including vegetation parameters from CalVeg vegetation parameters from the United States Forest Service.

Acquiring individual tree biomass from small-footprint lidar data might be considered a better method of estimating biomass as opposed to plot-level regression methods from ground-truth data or lidar data (Popescu, *et al.*, 2007). Although, with uncertainty in global positioning systems (GPS), ground-truthed measurements and uncertainty in detection of individual trees provides a means of error to lidar derived individual tree measurements. The effects and quality of a certain interpolation methods from point to grid can aggregate biomass uncertainty in DEM and DSM surfaces as well (Smith, *et al.* 2004). Certain lidar derived individual tree biomass estimations may also provide a means of quantifying uncertainty to plot-level regression methods

from ground-truthed data or vice-versa. In this study, ground-truthed GPS points were manually corrected to their nearest-neighboring tree-top or Canopy Height Model (CHM) value.

It was out of the scope out of this study to estimate biomass using destructive sampling and regression. The main objective of this study was to quantify above ground biomass characteristics in the study site relative to the Sierra Nevada Adaptive Management Project as well as:

i. Use ground-truthed observations to validate individual tree detection from lidar data using individual tree segmentation from TreeVaW software (Popescu, *et al.* 2004).
ii. Explore the use of Jenkins *et al.* 2003 biomass equations with vegetation parameters acquired from CalVeg data for a better estimation of total plot-level biomass in comparison to using individual tree-based approach.
iii. Use lidar based multivariate regression approaches to compare individual tree based measurement from lidar and ground-based measurements.
iv. Access the overall quality of using CalVeg and TreeVaW as vegetation and individual tree-detection parameters to estimate biomass at the individual-tree level.

As shown in the ***appendix*** section of this thesis, ground-truth data collected in this study provides a validation to detected biophysical parameters extracted from lidar data, including a validation to software that provide algorithms in extracting tree-height parameters such as TreeVaW. Since previous studies prove that tree-dbh is the most reliable variable for estimating biomass (Crow, 1971; Schroeder *et al.*, 1997), *figure A-2* (*Descriptive statistics on tree-height and observed diameter at breast height (dbh) of tree species from study area*) and *table A-1*(*Descriptive Statistics on ground inventory data including biomass based on Jenkins, et al. 2003*) describes the ground-inventory statistics for the area used in this study.

4.2 – Methods

Since tree-species are a component in measuring biomass based on equations in forestry management such as (Jenkins *et al.* 2003), the CalVeg dataset from US Forest Service was used to apply tree-species to our individual detected trees extracted from lidar-data. CalVeg datasets provide existing GIS vegetation maps that meet regional and national vegetation mapping standards. The methodology used to capture forest vegetation characteristics using automated includes methods such as remote sensing classification, photo-editing and field based oberservations (CalVeg Existing Vegetation, 2011). To read more on how CalVeg classifies vegetation parameters based on the level of National Vegetation Classification Standard hierarchy please see the following reference *CalVeg Existing Vegetation, 2011.* Although, this dataset is very coarse especially at the individual tree-level. Finer resolution datasets at an individual tree level do not yet exist.

Since this paper explores the variation of multiple methods of extracting biomass estimates from a forested region in the Sierra Nevada, a lidar point cloud extraction was performed on height-percentiles from raw lidar data. The raw lidar data is compressed into *.las* format, and it is necessary to extract the points from the raw lidar based on the bytes classified from TerraSolid's TerraScan software. In this case, python 2.6 was used to extract classified ground-points and vegetation points from the raw-lidar data. After both types of classification points were filtered (removal of outlying values in each point-dataset), each point dataset was interpolated using Universal Kriging at *1m* spatial-grid to create DEM, DSM and CHM products. The following information was extracted from the raw-lidar data including ground-truth data for each plot measured in the study area: height-percentiles based on discrete lidar points of plot (min, 1%, 5%, 10%, 25%...99%), Max, Mean, Standard Deviation, Coefficient of Variation.

Ground-truth data parameters extracted per plot in ground-truth data include: height of tree average, dbh average, and bulk data density averages to perform the multivariate regression to extract percentile heights from the raw lidar data. The assumption behind the multivariate regression technique from raw lidar to ground-truthed biomass estimates is that vegetative height in each plot extracted from lidar data has a independent relationship with dependent biomass estimates from ground-based inventory data (*table A-1)*. Refer to *figure 19* for percentile heights extract for specific ground-truth plots.

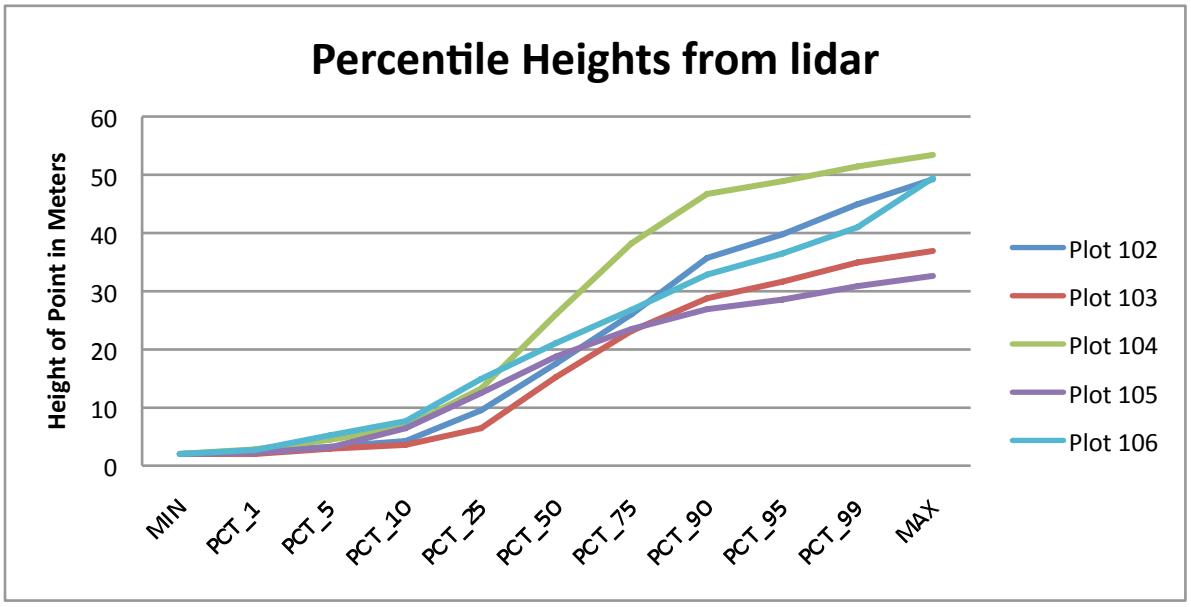


Figure 19: Percentile Height Information extracted from raw lidar data including a sub-set of plots in study area.

The software used for detecting individual trees from the CHM (interpolated tree-height surface) is implemented in TreeVaW software described (S.C. Popescu and R.H. Wynne, 2004; S.C. Popescu, R.H. Wynne and J.A. Scrivani, 2004). TreeVaW essentially executes an adaptive technique for local maxima focal filtering on a CHM surface. The study-area CHM was split into multiple subsections in ArcGIS 9.3, processed using ENVI software standard format, and then passed into the TreeVaW application. The result provided parameters such as Longitude

and Latitude in decimal degrees, Tree-Height and Crown-Radii in ASCII format. This TreeVaW software output was then imported back into ArcGIS for analysis to combine tree-height information with existing vegetation information for the study site. Refer to *figure 20* below for a complete flowchart of data processing from raw-lidar to the comparison of biomass estimates.

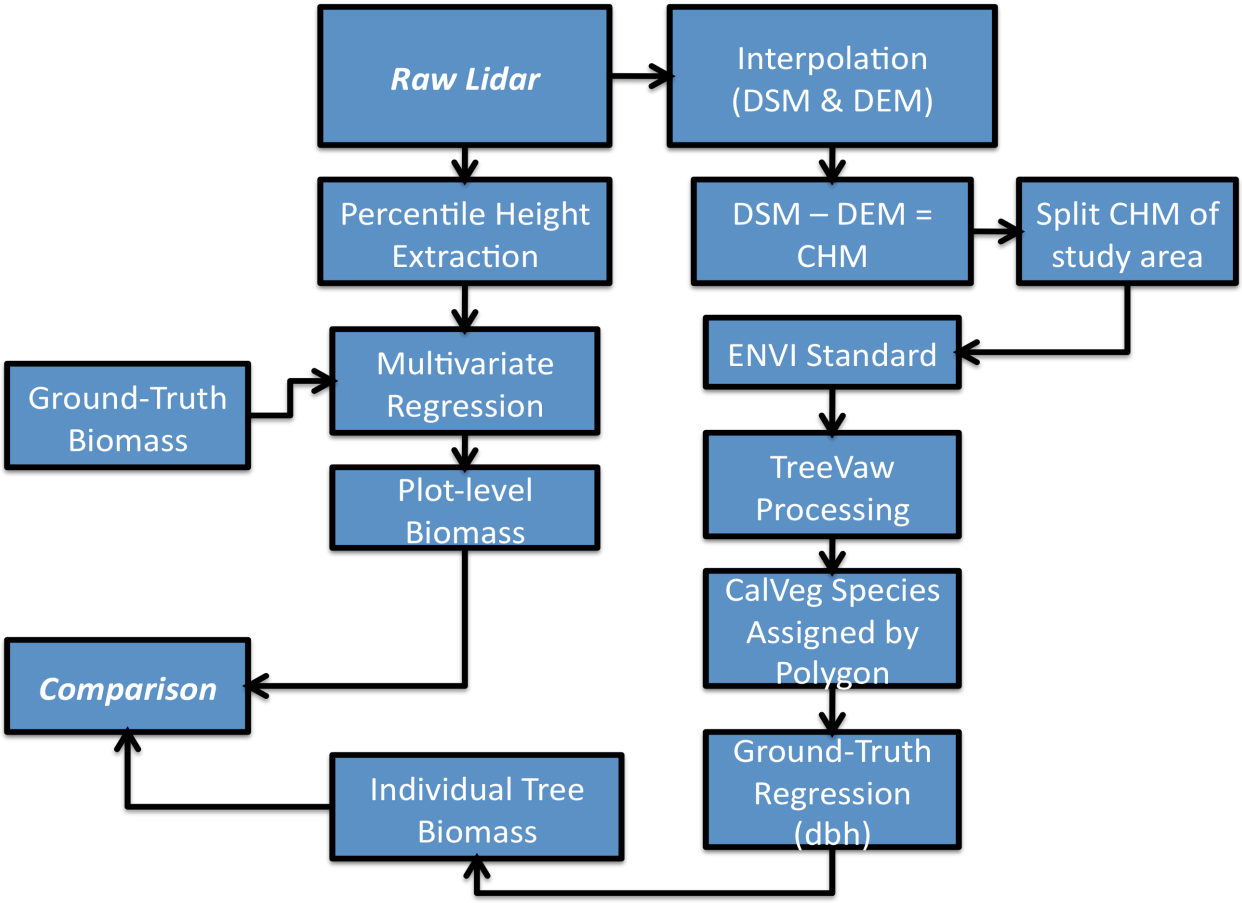


Figure 20: Flowchart of extraction of biomass estimates for the Sierra Nevada study area as described in *section 1.1*.

After importing the resulting individual tree-detection from TreeVaW into ArcGIS, a spatial join was then used to merge the existing vegetation type from CalVeg to the detected trees. A spatial join was performed in ArcGIS; each tree-height point acquired from TreeVaW that falls within that particular vegetation polygon is assigned that particular vegetation type from CalVeg. After this was performed for each sub-section of the study site, the final combined shapefile was then exported into ASCII format for analysis. Python 2.6 was used to

programmatically assign the biomass equations at the individual tree-level. A linear regression approach was used to assign dbh to a particular tree-species from the ground-based measurements (see ***appendix)*** to tree species classified and tree-heights detected from CalVeg and TreeVaw respectively. A buffer was used to acquire the number of detected, biomass, and statistics of individual trees for each plot to compare tree-heights assigned from TreeVaW to ground-based measurements (refer to *table 4* for results).

4.3 – Results

Biomass estimations were calculated and compared at the plot-level using three different methods: individual tree based approach, ground-truth approach, and multivariate regression lidar approach using percentile heights extracted from raw-lidar points. On average TreeVaW under-detected the amount of trees in comparison to ground-truthed data, based on 116 plot measurements. Due to the TreeVaW software under-detection, this also led to under-estimation of the amount of biomass at the plot-level. Since a variety of studies have included the use of TreeVaW with respect to biomass excluding study area regions in the Sierra Nevada (Popescu, *et al.* 2004; Popescu, *et al.* 2005; Lefsky *et al.* 2002), this might suggest that this software might not be as reliable on high dense canopy cover and vegetative regions.

Based on the multivariate regression approach on 9 different percentiles (1%, 5%, 10%, 25%, 75%, 90%, 95%, and 99%) point-cloud heights extracted from the raw-lidar for each plot in comparison to the dependent variable (ground-measured biomass) in this study, we found that the correlation r^2 value was *0.75*. This value can be a means of stating that solely using percentile heights in lidar data we are able to describe a relatively high correlation of biomass using species, dbh and tree-height information from ground-truth data at the plot-level. Since

tree-dbh was found to be the most reliable method of describing biomass, it should also be noted that tree-height information is also a major factor in estimating biomass at both the plot-level and individual tree-level. This multivariate lidar regression approach seems to be a reliable method of calculating biomass for an entire study-site rather than the individual tree-based approach using TreeVaW and CalVeg since TreeVaW on average under-detects the amount of trees within a certain plot, there is a great underestimation of biomass at the plot-level. The *discussion* section in this chapter will provide further insight into these using these particular methods for estimating biomass for particular study-areas and situations.

Table 4: Descriptive Statistics on individual tree-biomass, results based on average for 115 plots. Including r^2 values between programmatic TreeVaW/Calveg approach and Ground-based approach.

Data	Biomass(kg)	Tree Count	DBH(cm)	Tree-Height(m)
TreeVaW	7739	10.55	40.46	17.50
Ground Truth	13557	13.03	39.68	18.50
r^2	0.31	0.03	0.28	0.42

Since each and every tree was measured within the 12.62 m radius (distance measured from chosen plot-center) plot in the ground-inventory data, it was acceptable to compare measured ground-based tree-count with tree-count of TreeVaW (individual tree based-approach). For *116* measured plots, TreeVaW detected an average of *10* trees per plot. Ground-based tree-count measured an average of *13* trees per plot. TreeVaW under-detected tree-heights from ground-truthed data; this suggestion might also include lidar-data underestimates due to first-return hits in the lidar data (Popescu *et al.*, 2002; Juan C. Suarez *et al.* 2005). This underestimate of biomass at the invidividual tree level in comparison to plot-level is shown in *figure 21* as well as biomass estimates from the study area in the Sierra Nevada in *figure 22.*

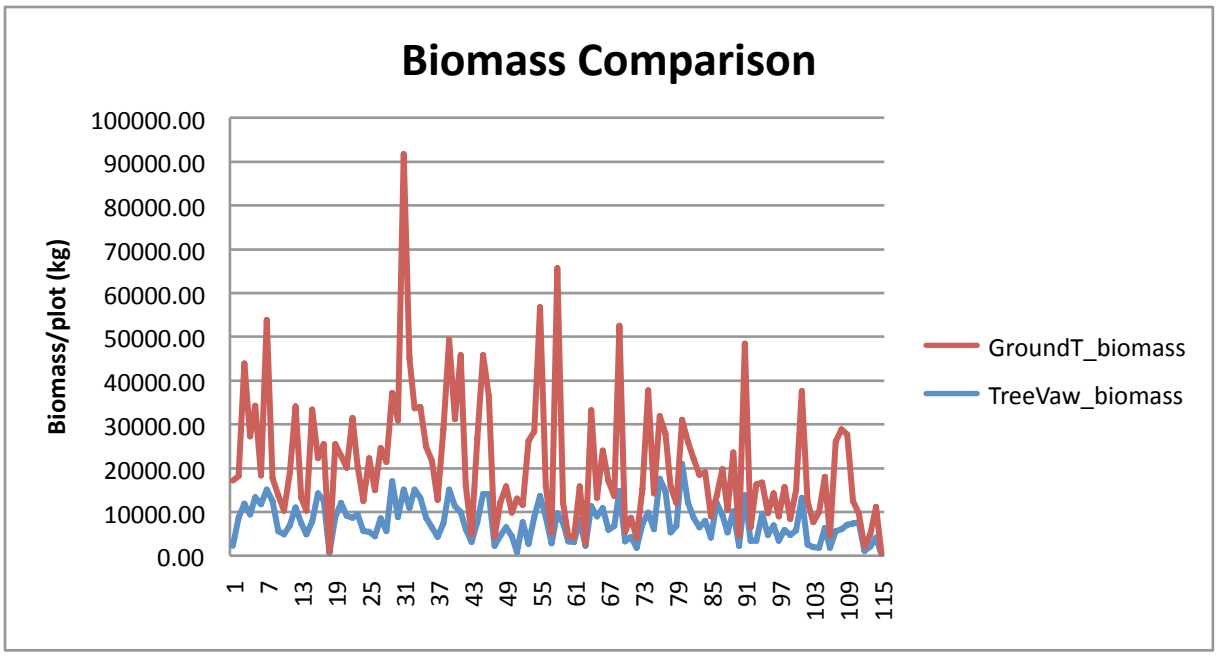


Figure 21: Individual tree (blue-line) and ground-based biomass comparison (red-line). Units are in kilograms/plot.

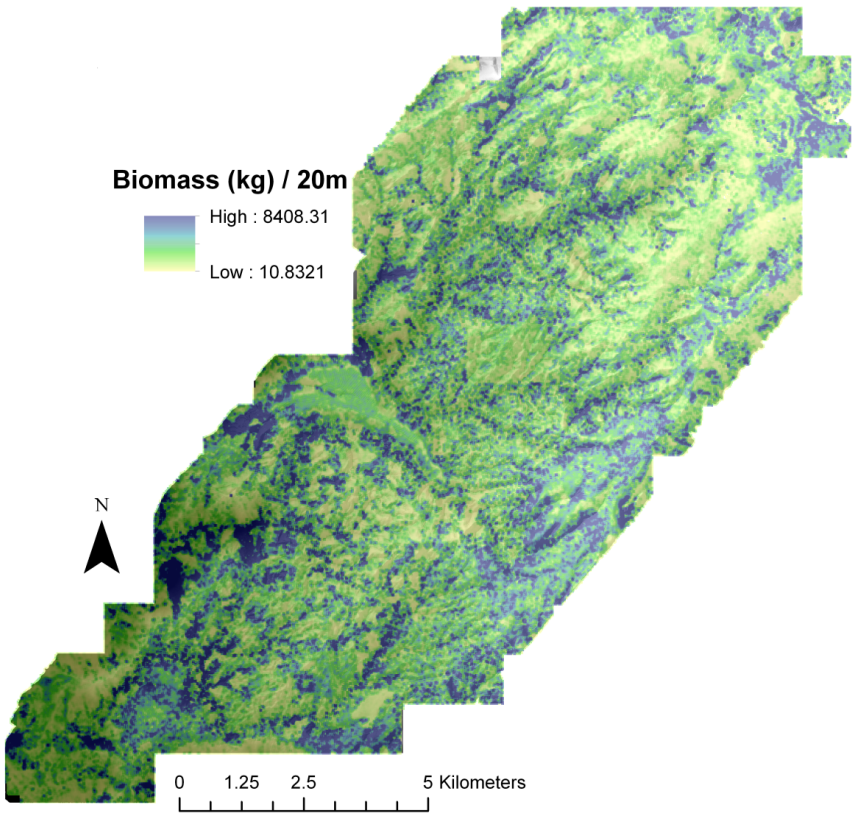


Figure 22: Biomass estimates for study area using TreeVaW tree-detection, CalVeg vegetation, & ground-truth regression. Biomass is measured in kilograms/20m^2.

4.4 – Discussion

Currently there exists a variety of studies that suggest discrete lidar has become a proven technology to estimate forest biophysical parameters automatically at both the plot and stand level (Popescu *et al.* 2007; Lefsky *et al.* 2002), although not a variety of studies have researched the comparison of extracting biomass from a variety of approaches from lidar data with the usage of existing vegetation layers or the comparison thereof. The unique contribution and suggestion of this paper is that users of TreeVaW and any other individual tree-detection algorithm software, vegetation classification and other schemas should thoroughly compared and validated before their use especially for high dense canopy cover and highly vegetated regions with a very thorough ground-based inventory dataset as described in the ***appendix***. Topographic variability including the coefficient of variation in the CHM surface was also computed for plot-level biomass comparisons where $Z^{predicted}$ is the interpolated canopy height surface and Z^{real} is the actual lidar point at i, Z is the mean elevation and n is the number of lidar points within the plot specific 12.62 m radius.

$$CV = \sqrt{\frac{\sum_{i=1}^{n}\left(Z_i^{predicted} - Z_i^{real}\right)^2}{n}} / Z \qquad (23)$$

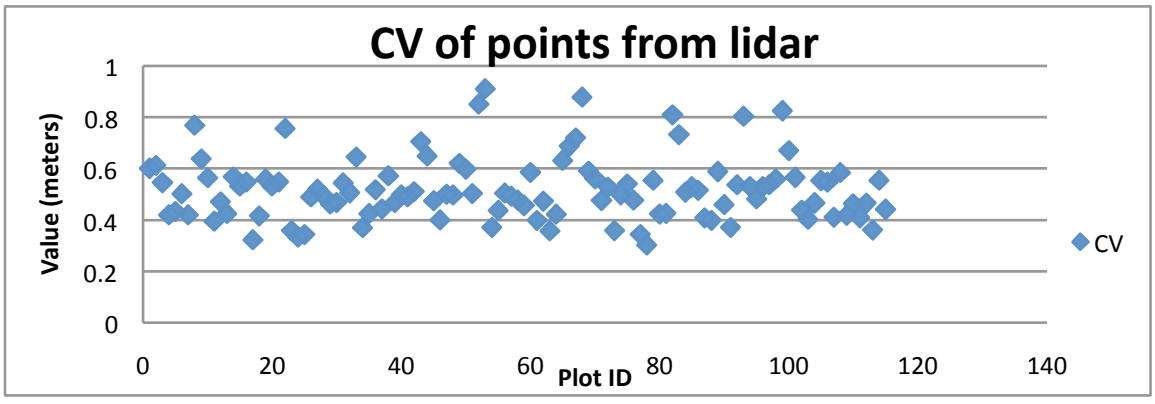


Figure 24: Coefficient of variation of raw lidar data for study for 116 measured plots.

Topographic variability as shown in equation (*23*) and *figure 24* for our study area explains terrain complexities within vegetation surfaces within each plot. This variable may be explored in created Canopy Height Models from a variety of interpolation techniques include canopy coverage, lidar-point data density and multiple individual tree extraction techniques. Sensitivity analysis and error analysis between products aggregates when there is error at the beginning stages in spatial processes and should be noted when trying to extract particular information especially from that of lidar data.

Individual tree extraction or identification methods along with the classification of tree-species in this study is lacking due to the availability of remotely sensed methods and data. Although some studies focus on other methods of deriving biomass such as scale-invariant approaches and the fusion of multi-spectral imaging (Zhao *et al.* 2009, Popescu *et al.* 2004). Other approaches of acquiring biomass use training data at the individual tree-level have been studied (Zachary *et al.* 2005) as well as a comparison of biomass across multiple communities characterized by distinctive tree species (Lefsky, *et al.* 2002). A comprehensive comparison of each particular method used to extract biomass information from lidar data from both the individual tree and plot-level would be an interesting study along with different biomes. Another interesting study would be the comparison of multiple individual tree extraction methods from canopy interpolated surfaces or directly from the lidar point-cloud as suggested in these studies (Daniel A. Zimble *et al.* 2003; Juan C. Suarez *et al.* 2005; Li, W. *et al.* 2011). Although this study aims to provide suggestions into what methods to extract biomass from lidar data using a multivariate regression approach and ground-truth approach, that are the most used in practice and are the simplest forms in extracting biomass estimates from lidar. Individual tree extraction and species classification are two underlying issues in detecting biomass, as well as a few

unanswered questions that needed to be answered in this study with regard to error propagation with respect to ground-truth data, available equations and vegetation datasets:

1) How to apply biomass equations to CalVeg species misclassification (*i.e* CalVeg misclassified a tree as water)?
2) How to apply ground-truth based regression to species classification outside of ground-inventory (*i.e* tree-species detected in CalVeg doesn't exist in ground-inventory data or Jenkins *et al.* 2003 biomass equations)?
3) What is the uncertainty in Jenkins *et al.* 2003 biomass equations?
4) How reliable are GPS recorded ground-based measurements? Can we derive a means of error from lidar detected trees visually?

A further in-depth analysis of acquiring biomass and the comparison of approaches is needed in order to further quantify which means is appropriate for a particular study area including better methods of extracting individual trees from lidar as well as species classification using multi/hyperspectral imagery. The individual tree extraction method along with species classification uncertainty is also needed to acquire error bounds in biomass estimates. Also, since lidar datasets tend to be relatively large and very time-consuming to process, computationally fast methods are needed in order to compute biomass in a timely manner for large spatial regions.

Ground-truth observations should be measured carefully since there are inherit errors in GPS and Vertex Ultrasonic Hypsometer to measure tree-heights. In this study, we assumed the GPS of the lidar was more accurate due to its mounting aircraft and full-view of sky. Since each tree detected that was ground-based had some inherit error due to GPS being inhibited by the

canopy, each ground-measured tree was moved to its nearest neighboring, and relatively closest CHM pixel tree-top value. A comparison of GPS error between manually corrected ground-based measurements and raw ground-based measured can yield error bounds of GPS devices. A comparison between Vertex Ultrasonic Hypsometer and the CHM acquired from interpolated lidar data can also be performed to access the validity in the Hypsometer device. Also since the interpolation between point-to-grid is subject to much uncertainty, it should be noted when creating a CHM to use a higher resolution (*0.5m preferably*) along with a reliable interpolation method for the study area of interest. Descriptive statistics on ground-truthed data is available in the ***appendix*** section of this thesis.

5 – Web-based Digital Library Development

5.1 - Introduction

As the cyber infrastructure supporting environmental observations expands, managing, sharing, and extracting information from data, many of which are continuously changing, is becoming an increasingly challenging problem. Flexible data repositories are needed to manage heterogeneous data and metadata streams, ranging from large spatial data sets (e.g., remote sensing products) to *in situ* sensor and sensor web time series, to results from manual sampling campaigns. In particular, environmental datasets pose significant technical challenges in terms of management and web-based storage and retrieval of these datasets. To promote data sharing and maximize information extraction, these repositories need to be accessible to a broad spectrum of users with an equally broad range of familiarity with information management and database querying skills. This paper provides a framework of developing a web-based storage and retrieval system to manage, store and share relatively large amounts of spatial data and other datasets using a lightweight, user-friendly open-source content management system using an object-oriented database and scripting languages. The combination of Zope, Python and the Google Maps Application Programming Interface (API) provides a novel user-friendly and relatively quick method for parsing, sharing and visualizing large spatial datasets. Eventual optimizations to the system will lead to even faster retrieval and easy distribution of this packaged software to run on a variety of platforms both UNIX and Windows based.

Earth systems observational capabilities have been increasing rapidly over the past few decades along with the increasing need for a multidisciplinary perspective in terms of data management and analysis. Data are increasingly available at a variety of spatial and temporal scales from remote sensing products (e.g., high resolution images, multi- and hyperspectral

products, lidar, etc.) and reliable *in situ* sensor platforms for monitoring meteorology, air quality, hydrology, water quality, and terrestrial and aquatic ecology. While the availability of such data can enable researchers to pursue complex lines of inquiry, rapid progress hinges on the need for easy access to and integration of these data. Hence, organizing heterogeneous datasets in a manner that facilitates shared access and analysis by a broad and interdisciplinary user base is an important task. This organization must encompass robust metadata schemes that enable specification of the datasets, including where, when and how they were obtained, as well as provenance information on newer versions of previous datasets in the event modifications are made.

Digital libraries (DLs) include a variety of digital content as well as the aggregation of multiple collections of metadata describing it (Baldonado *et al.* 1997). Libraries, museums, and universities have been rapidly moving toward a DL format, but find difficulties in building these services because of metadata quality and shareability issues such as: inconsistencies in metadata, too much technical information, lack of key contextual information and lack of conformance of technical standards (Shreeves *et al.* 2006). Analogously, environmental data streams are becoming increasingly digital in nature (Pundt and Bishr, 2002), yet possess a wide range of metadata attributes associated with the time, location, and method associated with their acquisition. Thus, given ease of use and familiarity of a DL format to most potential users, it is reasonable to suggest that DL systems combining web-based storage and retrieval of Earth systems science data and metadata would be more likely to attract a broad user base than, say, comparable database programs. This is significant, as the use of spatial databases and visualization of environmental datasets therein has also seen dramatic increase (Oosterom, and Lemmen, 2001).

Web-based information storage and retrieval systems are now offered by a variety of organizations, usually governmental, and include soil attributes, landcover, digital elevation models (DEMs), light detection and ranging (lidar) data, and others (USGS 2004; USDA 2011; NASA 2011). Although the majority of these systems are successful to some extent, access to centralized spatial data with descriptive metadata is often limited, and can be relatively difficult for a non-expert end-user to acquire and process. An example of a fast and user-friendly application is the Soil-Web product (Beaudette and O'Geen, 2009), an online soil survey that depicts a seamless coverage of soils information for California, Nevada and Arizona. Although queries for larger datasets are limited in the Soil-Web are limited, this product is a step forward in combining a variety of data into one centralized product using open-source applications. Another environmental application is a portal offering enhanced access to high-resolution topography data (V. Nandigam *et al.* 2010). These portals have become increasingly popular within the last decade, as web-based Application Programming Interfaces (APIs) and relational databases become increasingly easier to use. Although large volume environmental datasets, such as lidar, make it difficult for a typical user to process this type of data, supercomputer clusters and the use of parallel processing can decrease processing times exponentially. The need for centralization of data and a simple way to share data along with descriptive metadata is essential, but lacking in this era of web-based technology.

This paper provides a solution to the challenging problem of storing and sharing heterogeneous spatial and temporal environmental datasets using an open source, lightweight, user-friendly web-based digital library (DL). The framework presented is novel in that it is particularly adept for managing environmental datasets, non-spatial, spatial and temporal. The following sections provide an overview of the component software and how it was used to

develop this DL. This is followed by a description of several cases studies in which the DL is being implemented to manage data and facilitate collaboration in both research and educational contexts.

5.2 - Methods

The open-source content management system used in this framework is Zope 3 and programming language Python 2.5. Zope standalone provides security as well as user-defined access control to particular folders or sites created, and can provide a quick and simple solution to the use of storing and sharing data. Zope is an easily installable content management system that is primarily python extensible, security assignable, and works under a relational database or "catalog". Zope also supports many different web-based scripting languages such as Hyper-Text Markup Language (HTML), JavaScript, Cascading Style Sheets (CSS), etc. These main tasks were to be accomplished to have a successful working web-based digital library for environmental datasets. Some of the characteristics of the Zope-based system are:

i. Fast and easy way to store and share data.

ii. Secure way to store and share data (assignable permissions via access control).

iii. Metadata functionality for each item or list of items (multi-file control).

iv. Assignable metadata for each item or list of items uploaded.

v. Assignable spatial extents to be displayed and queried via dragZoom feature.

vi. File Transfer Protocol (FTP) for large spatial files.

vii. Structured Query Language (SQL) extensible.

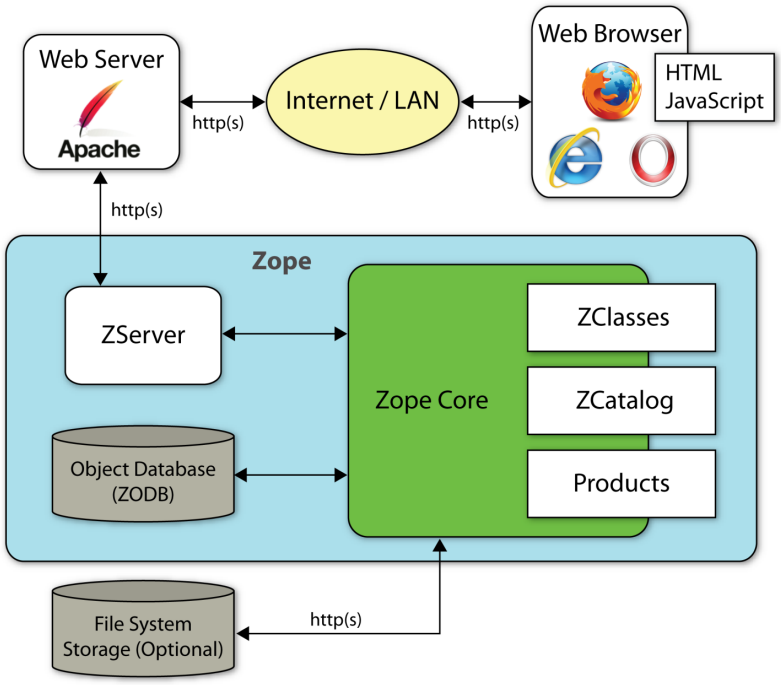


Figure 25: Zope interaction between ZServer, Apache and Web-browser with extensions such as Products and File System Storage.

Access control can be defined in the Security and local roles section of Zope. In this instance, our site is connected via Lightweight Directory Access Protocol (LDAP) for users and is extensible through Zope. Those outside of LDAP system still may have access privilege to certain content within the site. In this case, we have designated specific teams allowed to access only their folders data. For a descriptive overview on setting access control privileges visit zope.org.

All data items that are stored into our digital library are controlled by a ZCatalog for quick indexing, searching and queries. A ZCatalog is a Zope object that can be added to a Folder in the site, managed through the web and extended in many ways. It is also very simple to create search forms and report results from queries using this object. To allow users to search for individual words within the description, metadata can be assigned to each particular file. In this case, we present a method for a set of vocabularies used in SNAMP and NCZO projects

(described in our case studies) to define spatial metadata and a programmatic approach to assigning this metadata along with querying the metadata.

The front-end of our current digital library system's graphical user interface (GUI) is controlled by an HTML file, which has preloaded CSS and JavaScript. The current setup for the GUI mimics Zope's file storage structure with the ability to add, delete, copy, files from a user. Since Zope only allows the control of single file upload per-submission, we can also have created another form for batch uploads. We also encourage users to add zip files for large amount of files of the same type.

Another unique aspect of our digital library is that a separate python script with embedded HTML in Zope controls the metadata structure for each particular file. This metadata assignment can be entered manually via the python form or programmatically parsed from an eXtensible Markup Language (XML). This flexibility allows users to assign multiple files the same metadata if needed. If the metadata field needs to be updated, we add this new attribute to the index of the ZCatalog.

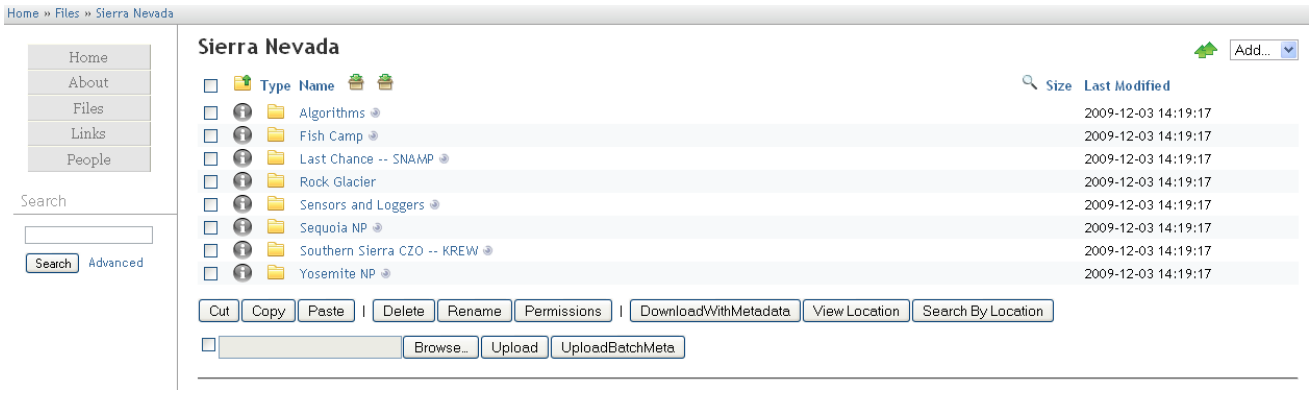


Figure 26: The GUI for our Digital Library using backend Zope functionality.

The ability to download multiple files along with assigned metadata (if also multiple) is another important feature that was added to this digital library. For example, if two separate types of spatial data types were added to the digital library along with two separate metadata

attributes within the same folder; a user should be able to download these separate data types along with assigned metadata. A “Download All” and “Download All with Metadata” functions have been implemented at the folder level. These functions return a zip file to the user after the desired files are selected for download. The zip file contains the metadata in XML format along with the specific file in its original data format. An FTP site for large data transfers can also be extended through Zope. In the spatial data case, we use FTP sites for large volume spatial data transfers as well as remote sensing images.

This digital library can host a variety of datasets including non-spatial data such as documents, images, etc. as well as temporal data, although, this digital library was built on the basis to support and handle spatial data files and formats. Metadata within the system although specifically formatted for environmental datasets can be generic, meaning the metadata schema in this system is user-defined. Depending on the amount of storage space allocated within the platform that is being used, this digital library can store and share this data. To query spatial data after the metadata form is filled (specifically the extent attributes), the Google Maps API displays those specific longitude and latitude attributes for selection. Also the ability of KML (Keyhole Markup Language) overlays are also able to be parsed into the map and downloaded. Point or polygon data is able to be queried from this map after it is loaded via click or dragZoom. The dragZoom feature is a feature that is built in JavaScript although called through the Google Maps API so that a user can select an area defined by that person’s bounding box extent. After the bounding box extent is drawn by the user, the resulting data is displayed in a website allowing the user to download the data along with its corresponding metadata in zip format. After the dragZoom bounding box is drawn, the resulting coordinates are passed back to the ZCatalog, returning any files whose spatial extent attributes fall within the region. If an

overlapping region is selected, a multi-tabbed infoWindow is generated to display a selection of data within the region of interest. If a user also clicks on the point or polygon displayed inside of the map, that user will be able to download that corresponding areas data with metadata within the Google Maps API infoWindow. Temporal data downloads are also able to be queried within the Google Maps API.

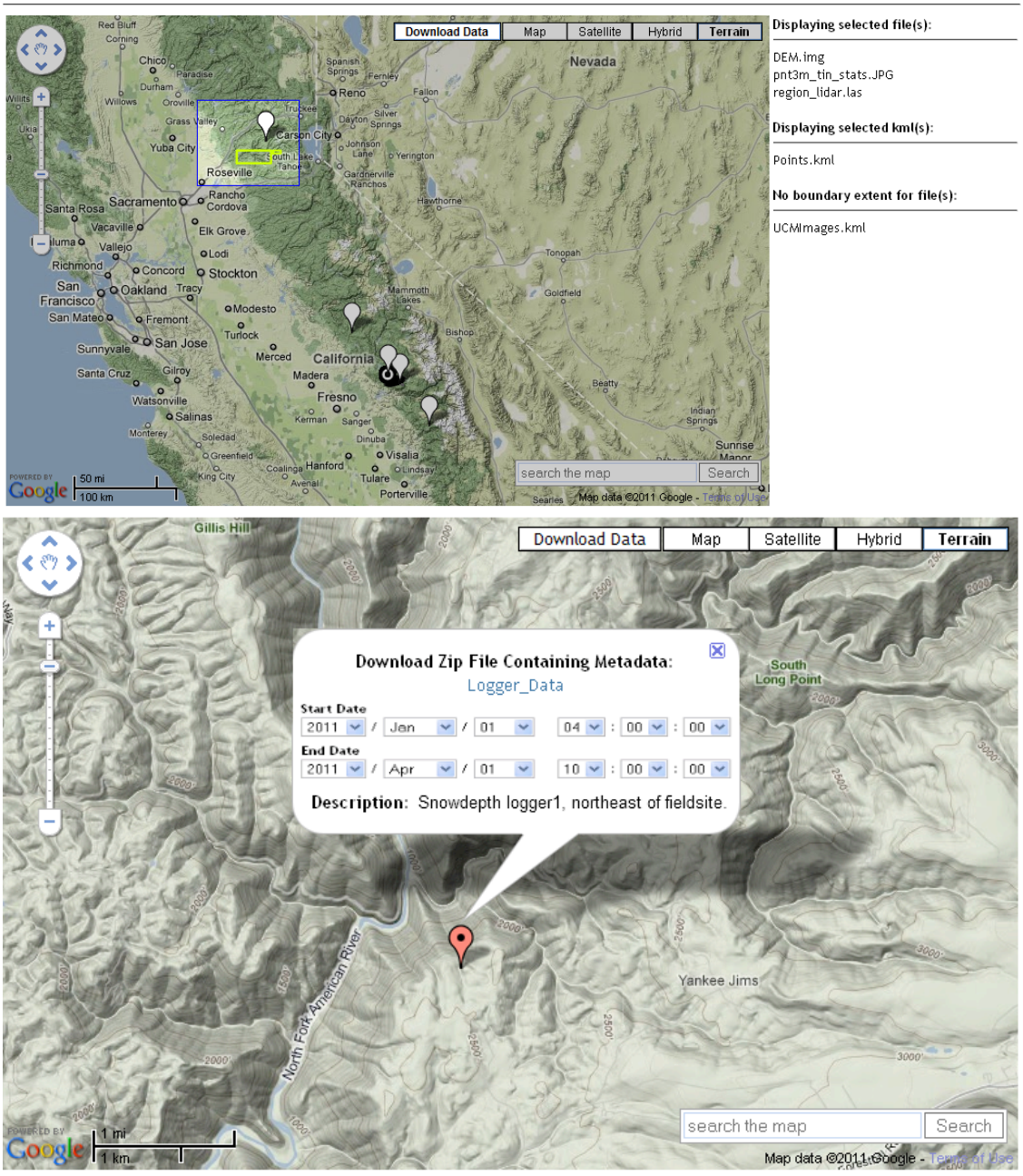


Figure 27: Google Maps API visualization of spatial data collected in the Sierra Nevada parsed from spatial extent attributes in metadata including Lidar and temporal data.

5.3 – Case Studies, Discussion and Acknowledgements

The earliest version of this DL approach was the Sierra Nevada San Joaquin Hydrologic Observatory (SNSJHO 2007). The SNSJHO was initially developed as a data repository for the U.S. National Science Foundation (NSF) WATERS Network California test bed (Montgomery *et al.*, 2007), The SNSJHO now serves as a data and information repository for multiple major research projects, including the SNAMP project (SNAMP, 2010), a joint effort by the University of California, state and federal agencies, and the public. The SNAMP project has been formed to develop, implement and test Adaptive Management processes through testing the efficacy of Strategically Placed Landscape Treatments (SPLATs) across four response variables, including: public participation, wildlife focusing on the Pacific Fisher and the California Spotted owl, water, along with fire and forest health (SNAMP, 2010). This diverse set of variables creates a broad range of data and information types. The SNSJHO DL also houses data emanating from the National Critical Zone Observatory (NCZO), a watershed-scale program that primarily investigates the processes that occur at and near the Earth's surface that are affected by fresh water, both are in need of an accessible system to store, share and retrieve spatial data along with metadata attributes (NCZO, 2010). Both the SNAMP and NCZO projects have implemented the DL to securely store, share and retrieve data ranging from documents to high volume spatial data. This online content management system is accessed daily by members including the public interested in acquiring spatial datasets and metadata attributes. A second DL example supports an education and outreach effort aimed at middle school students.

The files placed into our digital library then ultimately in the ZCatalog are considered Zope objects. Handling Zope objects solely is a limitation instead of handling the files themselves directly. This setback led us in the direction of pursuing a Zope and PostgreSQL

database connection instead of using the ZCatalog for file handling. The flexibility of a standalone PostgreSQL database allows for the centralized data access, essentially one database can host a large amount of spatial data, while our digital libraries can display and submit queries directly to the database. The current system in place won't allow for subsections of spatial data to be parsed, similar to (V. Nandigam *et al.* 2010) system. Another addition to the system would be to handle each spatial data by coordinates in a database. This type of setup would be ideal for those looking for a lightweight system to store and share data amongst a variety of users. Download speeds and transfer rates would be dependent on user's bandwidth.

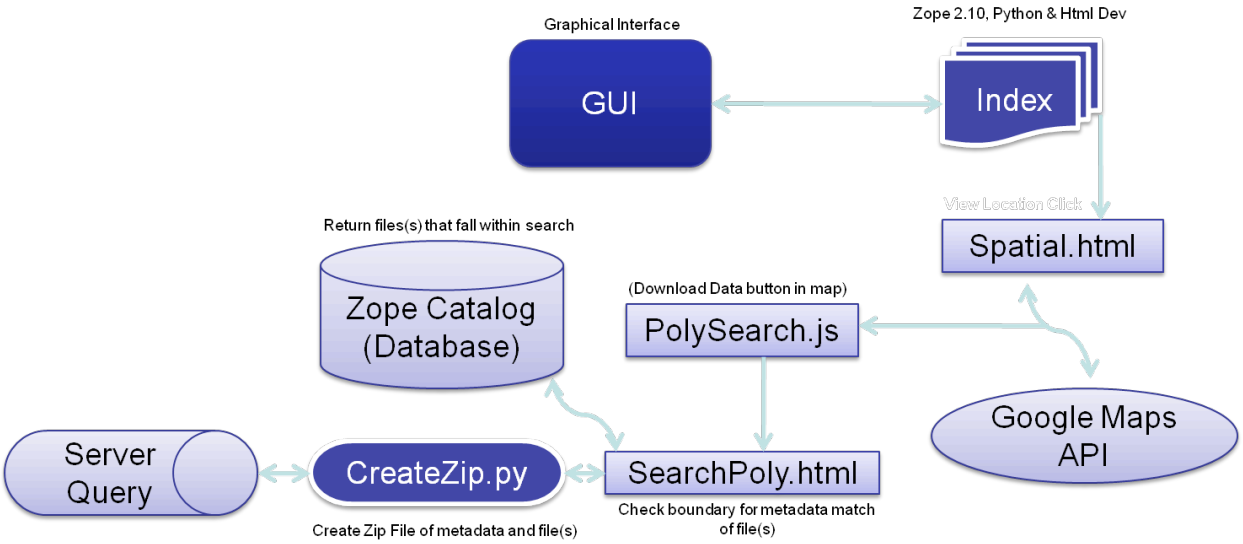


Figure 28: Flow of data execution when queried from Google Maps API

I'd like to acknowledge the developer who created this first generation of this software system Jason Fisher, along with those who aided in the recent developments of this digital library in order of contribution: Otto Alvarez, Jacob Flanagan and Andrew Zumkehr. Also, the Sierra Nevada Research Institute and the National Science Foundation for help in funding this project.

References

(2011). "Earth Science Data and Services Directory: Global Change Master Directory Web Site." from http://gcmd.nasa.gov/.

(2011). "LANDFIRE Data Products. *LANDFIRE Homepage*.", from http://www.landfire.gov/vegetation.php.

(2011). "NSF OpenTopography Facility | Home. San Diego Supercomputer Center."

Achaichia, B. A. S.-O. a. N. (2004). "Measuring forest canopy height using a combination of lidar and aerial photography." International Archives of Photogrammetry and Remote Sensing 35(4): 22-24.

Aguilar, F. J., F. Aguera, et al. (2005). "Effects of terrain morphology, sampling density, and interpolation methods on grid DEM accuracy." Photogrammetric Engineering and Remote Sensing 71(7): 805-816.

Armstrong, M. (1984). "Problems with Universal Kriging." Journal of the International Association for Mathematical Geology 16(1): 101-108.

Baldonado, M., C.-C. K. Chang, et al. (1997). "The Stanford Digital Library metadata architecture." International Journal on Digital Libraries 1(2): 108-121.

Bater, C. W. and N. C. Coops (2009). "Evaluating error associated with lidar-derived DEM interpolation." Computers & Geosciences 35(2): 289-300.

Beaudette, D. E. and A. T. O'Geen (2009). "Soil-Web: An online soil survey for California, Arizona, and Nevada." Computers & Geosciences 35(10): 2119-2128.

Bojanov, B. D., H.A. Hakopian, et al. (1993). "*Spline Function and Multivariate Interpolation*." Kluwer, Norwell, Massachusetts.

Bolstad, P. V. and T. Stowe (1994). "An Evaluation of Dem Accuracy - Elevation, Slope, and Aspect." Photogrammetric Engineering and Remote Sensing 60(11): 1327-1332.

Bortolot, Z. J. and R. H. Wynne (2005). "Estimating forest biomass using small footprint LiDAR data: An individual tree-based approach that incorporates training data." Isprs Journal of Photogrammetry and Remote Sensing 59(6): 342-360.

Brovelli, M. A., Cannata, M. and Longoni, U. M. (2004). "LIDAR Data Filtering and DTM Interpolation Within GRASS." Transactions in GIS 8: 155-174.

Callaway, R. M., E. H. Delucia, et al. (1994). "Biomass Allocation of Montane and Desert Ponderosa Pine - an Analog for Response to Climate-Change." Ecology 75(5): 1474-1481.

Caruso, C. and F. Quarta (1998). "Interpolation methods comparison." Computers & Mathematics with Applications 35(12): 109-126.

Chaplot, V., F. Darboux, et al. (2006). "Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density." Geomorphology 77(1-2): 126-141.

Coyeney, S., A. S. Fotheringham, et al. (2010). "Dual-scale validation of a medium-resolution coastal DEM with terrestrial LiDAR DSM and GPS." Computers & Geosciences 36(4): 489-499.

Cressie, N. (1988). "Spatial Prediction and Ordinary Kriging." Mathematical Geology 20(4): 405-421.

Cressie, N. (1990). "The Origins of Kriging." Mathematical Geology 22(3): 239-252.

Crow, T. R. (1971). "Estimation of biomass in an even-aged stand - Regression and “mean tree” techniques." In *Proceedings of the 15th IUFRO Congress, Section 25, Forest Biomass Studies*, 15-20 March 1971, Gainesville, Fla. University of Maine, Orono, Maine: 35-38.

David W. S. Wong, C. V. W. (1996). "Spatial Metadata and GIS for Decision Support,hicss." 29th Hawaii International Conference on System Sciences (HICSS), Collaboration Systems and Technology 3: 557.

Guo, Q. H., W. K. Li, et al. (2010). "Effects of Topographic Variability and Lidar Sampling Density on Several DEM Interpolation Methods." Photogrammetric Engineering and Remote Sensing 76(6): 701-712.

Hodgson, M. E. and P. Bresnahan (2004). "Accuracy of airborne lidar-derived elevation: Empirical assessment and error budget." Photogrammetric Engineering and Remote Sensing 70(3): 331-339.

Hodgson, M. E., J. R. Jensen, et al. (2003). "An evaluation of LIDAR- and IFSAR-derived digital elevation models in leaf-on conditions with USGS Level 1 and Level 2 DEMs." Remote Sensing of Environment 84(2): 295-308.

Hodgson, M. E., Jensen, J., Raber, G., Tullis, J.,Davis, B.A., Thompson, G. and Schuckman, K (2005). "An evaluation of LiDAR-derived elevation and terrain slope in leaf-off condition. ." Photogrammetric Engineering and Remote Sensing 71: 817-23.

Jackson, R. B., E. G. Jobbagy, et al. (2005). "Trading water for carbon with biological sequestration." Science 310(5756): 1944-1947.

Jenkins, J. C., D. C. Chojnacky, et al. (2003). "National-scale biomass estimators for United States tree species." Forest Science 49(1): 12-35.

Jensen, J. R. (2000). "Active and Passive Microwave, and LIDAR Remote Sensing." Remote Sensing of the Environment: An Earth Resource Perspective, Prentice Hall, New Jersey,: 285-332.

Kato, A., L. M. Moskal, et al. (2009). "Capturing tree crown formation through implicit surface reconstruction using airborne lidar data." Remote Sensing of Environment 113(6): 1148-1162.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model Selection." Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 14: 1137-1143.

Lal, R. (2004). "Soil carbon sequestration to mitigate climate change." Geoderma 123(1-2): 1-22.

Lefsky, M. A., W. B. Cohen, et al. (2002). "Lidar remote sensing of above-ground biomass in three biomes." Global Ecology and Biogeography 11(5): 393-399.

Li, W., Guo, Q., and Kelly, M (2011). "A new method for segmenting individual trees from the lidar point cloud." Photogrammetric Engineering and Remote Sensing

Lloyd, C. D. and P. M. Atkinson (2002). "Deriving DSMs from LiDAR data with kriging." International Journal of Remote Sensing 23(12): 2519-2524.

Lucas, R. M., Lee, A. C., Bunting, P. J. (2008). "Retrieving forest biomass through integration of CASI and LiDAR data." International Journal of Remote Sensing 29(5): 1553-1577.

Merced, U. (2007). "Sierra Nevada San Joaquin Hydrologic Observatory." from https://eng.ucmerced.edu/snsjho.

Merced, U. (2011). "Sierra Nevada Adaptive Management Project." from http://snamp.cnr.berkeley.edu/.

Merced, U. (2011). "Southern Sierra Critical Zone Observatory (CZO). Sierra Nevada Research Institute."

Montgomery, J. L. T., Harmon, B. Minsker, J. Schnoor, W. Kaiser, A. Sanderson, C.N. Haas, R. Hooper, N.L. Clesceri, W. Graham, and P. Brezonik (2007). "The WATERS Network: an integrated environmental observatory network for water research. ." Environmental Science and Technology 41(19): 6642-6647.

Moses Azong Cho, A. S., Fabio Corsi, Sipke E. van Wieren, Istiak Sobhan (2007). "Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression." International Journal of Applied Earth Observation and Geoinformation 9(4): 414-424.

P. J. M. van Oosterom, C. H. J. L. (2001). "Spatial data management on a very large cadastral database." Computers, Environment and Urban Systems 25(4-5): 509-528.

Polis, M. F., and D.M. McKeown, 1992 (1992). "TIN generation from digital evaluation models." Computer Vision and Pattern Recognition Proceedings CVPR 787-790.

Popescu, S. C. (2005). "Estimating biomass of individual pine trees using airborne lidar." Biomass and Bioenergy 31(9): 646-655.

Popescu SC, R. H. W., Ross F. Nelson (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 SC, W. R., and Scrivani JA. (2004). "Fusion of smallfootprint lidar and multispectral data to estimate plot-level volume and biomass in deciduous and pine forests in Virginia, USA." Forest Science 50: 551-65.

Popescu, S. C. and R. H. Wynne (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 and Remote Sensing 70(5): 589-604.

Priestnall, G., J. Jaafar, et al. (2000). "Extracting urban features from LiDAR digital surface models." Computers, Environment and Urban Systems 24(2): 65-78.

Pundt, H. and Y. Bishr (2002). "Domain ontologies for data sharing-an example from environmental monitoring using field GIS." Computers & Geosciences 28(1): 95-102.

Roberts, S. D., T. J. Dean, et al. (2005). "Estimating individual tree leaf area in loblolly pine plantations using LiDAR-derived measurements of height and crown dimensions." Forest Ecology and Management 213(1-3): 54-70.

Schroeder, P., S. Brown, et al. (1997). "Biomass estimation for temperate broadleaf forests of the United States using inventory data." Forest Science 42(3): 424-434.

Seamus Coveney, A. S. F., Martin Charlton, Timothy McCarthy (2010). "Dual-scale validation of a medium-resolution coastal DEM with terrestrial LiDAR DSM and GPS." Computers & Geosciences 36(4): 489-99.

Shreeves, S. L., Jenn Riley & Liz Milewicz (2006). "Creating Shareable Metadata.Web-Wise 2006 Pre-Conference. Los Angeles, CA."

Smith, S. L., Holland, D.A., Longley, P.A. (2004). "The importance of understanding error in LiDAR elevation models." 20th ISPRS conference: 488.

St-Onge, B., J. Jumelet, et al. (2004). "Measuring individual tree height using a combination of stereophotogrammetry and lidar." Canadian Journal of Forest Research 34(10): 2122-2130.

Suarez, J. C., C. Ontiveros, et al. (2005). "Use of airborne LiDAR and aerial photography in the estimation of individual tree heights in forestry." Computers & Geosciences 31(2): 253-262.

Thompson, J. A., J. C. Bell, et al. (2001). "Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling." Geoderma 100(1-2): 67-89.
Tobler, W. R. (1970). "Computer Movie Simulating Urban Growth in Detroit Region." Economic Geography 46(2): 234-240.
Toyra, J. and A. Pietroniro (2005). "Towards operational monitoring of a northern wetland using geomatics-based techniques." Remote Sensing of Environment 97(2): 174-191.
Treuhaft, R. N., G. P. Asner, et al. (2003). "Structure-based forest biomass from fusion of radar and hyperspectral observations." Geophysical Research Letters 30(9): -.
USDA-Forest Service, P. S. R., Remote Sensing Lab. (2004). "Existing Vegetation - CALVEG, [ESRI personal geodatabase]."
USDA (2011). "Web Soil Survey ".
USGS. (2004). "Light Detection and Ranging (LIDAR) Viewer Provides Free Online Data with NED, SRTM, Landsat, maps, orthoimagery, elevation ", from http://lidar.cr.usgs.gov/LIDAR_Viewer/viewer.php.
Viswanath Nandigam, C. B., and Christopher Crosby (2010). "Database Design for High-Resolution LIDAR Topography Data." Scientific and Statistical Database Management. Lecture Notes in Computer Science 6187: 151-159.
Yu, X. W., J. Hyyppa, et al. (2004). "Automatic detection of harvested trees and determination of forest growth using airborne laser scanning." Remote Sensing of Environment 90(4): 451-462.
Zachary J. Bortolot, R. H. W. (2005). "Estimating forest biomass using small footprint LiDAR data: An individual tree-based approach that incorporates training data." ISPRS Journal of Photogrammetry and Remote Sensing 59(6): 342-360.
Zhao, K. G., S. Popescu, et al. (2009). "Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers." Remote Sensing of Environment 113(1): 182-196.
Zimble, D. A., D. L. Evans, et al. (2003). "Characterizing vertical forest structure using small-footprint airborne LiDAR." Remote Sensing of Environment 87(2-3): 171-182.

Appendix

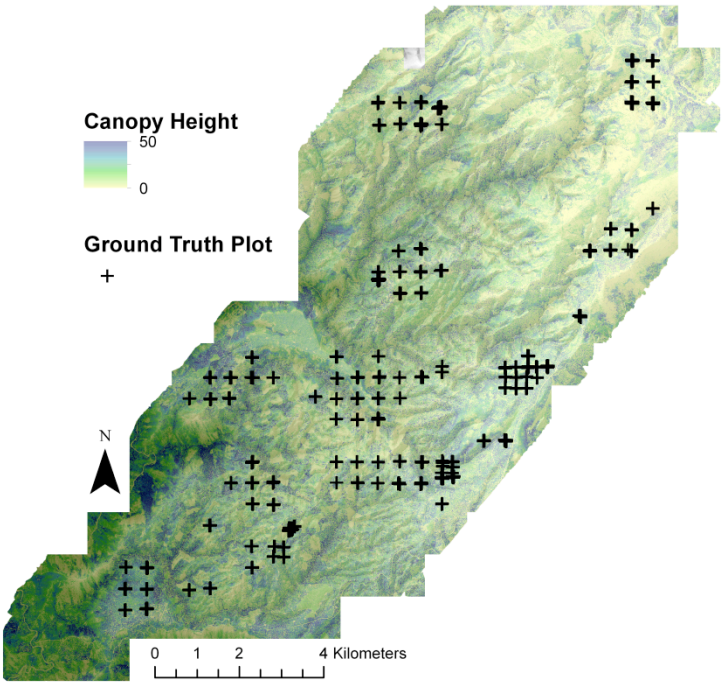


Figure A-1: Tree-heights of study area from CHM filtered at 50 meters, including ground-truthed plots.

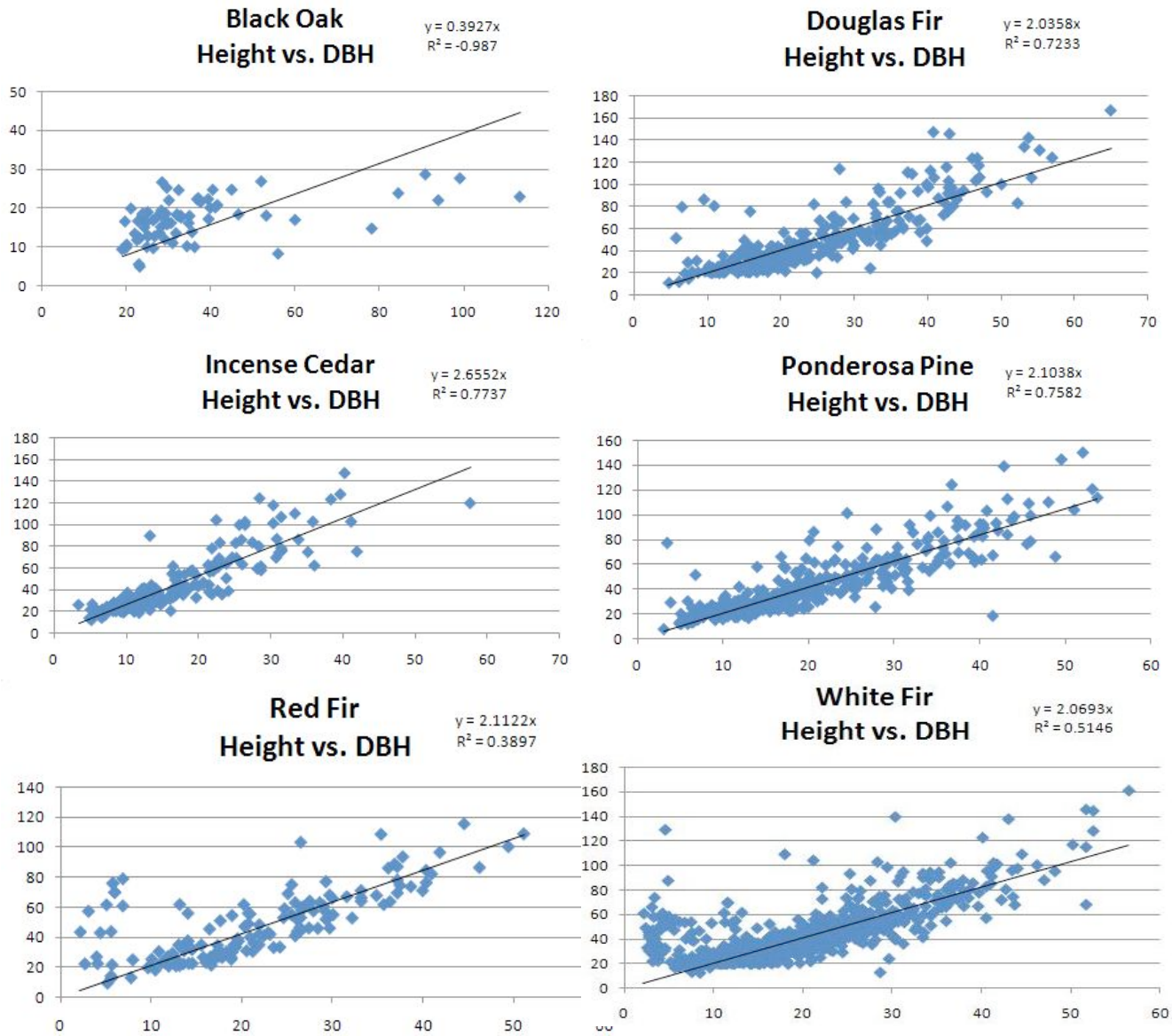


Figure A-2: Descriptive statistics on tree-height and observed diameter at breast height (dbh) of tree species from study area.

Table A-1: Descriptive Statistics on ground inventory data including biomass based on Jenkins, *et al.* 2003.

Species				
Red Fir		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	9.50	2.10	21.07
	Maximum	115.50	51.20	10367.59
	Range	106	49.10	1943.94
	Standard Deviation	22.95	10.91	10346.52
	Average	48.52	21.33	1715.92
White Fir		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	12.20	2.10	39.19
	Maximum	160.90	56.50	23601.25
	Range	148.70	54.40	23562.05
	Standard Deviation	20.87	9.11	2059.98
	Average	40.53	18.79	1193.10
Douglas Fir		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	10.8	4.7	36.01
	Maximum	167	65	29010.64
	Range	156.2	60.3	28974.62
	Standard Deviation	28.62	11.1	3833.24
	Average	48.03	24	2332.62
Pine (All)		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	7.5	3.1	10.70
	Maximum	150.6	53.7	15907.02
	Range	143.1	50.6	15896.32
	Standard Deviation	25.84	11.14	2057.23
	Average	42.47	19.95	1245.21
Incense Cedar		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	12.4	3.4	38.64
	Maximum	147	57.7	10309.13
	Range	134.6	54.3	10270.49
	Standard Deviation	26.85	8.91	1603.61
	Average	43.48	16.37	1034.62
Black Oak		**DBH (cm)**	**Tree-Height (m)**	**Biomass (kg)**
	Minimum	19	5.1	73.62
	Maximum	113.3	28.6	1540.08
	Range	94.3	23.5	1466.45
	Standard Deviation	19.64	5.25	284.20
	Average	36.25	16.98	256.51