Guidelines to estimate forest inventory parameters from lidar and field plot data

Companion document to the Advanced Lidar Applications--Forest Inventory Modeling class.

Authors and Contributors: Denise Laes, Steven E. Reutebuch, Robert J. McGaughey, Brent Mitchell
June, 2011.

Table of Contents

Overview

This document is intended to accompany the “Advanced Lidar Applications--Forest Inventory Modeling” training, however, it can also serve as a stand-alone reference or refresher for experienced users. Estimating forest inventory parameters from lidar and field plot data involves four major steps including: 1) collecting and preparing the forest inventory data, 2) preparing the lidar data, 3) Modeling (i.e., identifying and testing relationships between lidar derived variables and forest inventory variables), and, 4) Applying the modeled relationships across the landscape. There are four main sections to this document—corresponding to the four major steps above.

Background

Discrete lidar data continues to prove itself useful in many natural resource applications. However, while nearly all lidar data can be useful for some applications, not all lidar datasets are equal. Probably the most important single characteristic that determines the appropriate use of a lidar dataset is the mean number of pulses/m^2. For example, relatively low pulse-density data (0.5 to 1 pulse/ m^2) is typically only useful for bare earth or terrain models. Medium pulse-density (1-3 pulses/ m^2) data has the additional potential of providing canopy height models. Forest structure information however, requires relatively high pulse-density data (typically >= 3 pulses/ m^2). In addition, meaningful forest structure information from lidar data requires a significant investment in field plot inventory data (existing plot data is usually not adequate). It also requires that the general procedures of this document—including identifying and testing statistical relationships between lidar derived variables and forest inventory variables—are performed successfully. In other words, high-quality (high pulse density) lidar data alone are insufficient for deriving detailed forest structure[1] information across a landscape—additional significant investments in field data, data processing, and statistical modeling are also required. Without making the additional required investments, the extra cost of acquiring high-quality lidar data is wasted.

Field Plot Data -- collecting and preparing the forest inventory data

Collecting field data is required to quantify forest attributes from lidar data. A well designed field protocol, ensuring measurements needed to either calculate or model the attributes that will be estimated from the lidar data, is time well spent and will eliminate the need for subsequent field visits. In order to establish relationships between lidar data and forest inventory data, **the following characteristics of the forest inventory data are critical**:

- Location—plots should be measured to an accuracy of one meter or less.
- Timing—plots should be measured within one growing season of lidar acquisition.
- Size—plots should be large enough (> 1/10th acre) to minimize edge effect and characterize the vegetation. In addition, plots should have a fixed radius (rather than a variable radius as is common in Common Stand Exams (CSEs)).
- Biomass—all biomass contributing to lidar data pulse returns should be measured (e.g. not just the big trees).
- Samples—must have enough plots for statistical validity and the plots must cover the full range of variability of the measurement of interest.
- Consistency—in addition to a consistent and relatively large size, plots should represent single conditions—and collect the same data fields for each plot.

[1] Lidar data alone can supply canopy height and percent canopy cover, however, it cannot provide detailed inventory parameters such as quantitative estimates of biomass without associated field plots.

When resource managers learn that field data are still a requirement to generate forest estimates from the lidar data, a common response is to suggest the use of available field inventory data for the study area. These available field inventory data typically meet the original objectives for which they were designed; however, almost invariably each lacks at least one of the critical components listed above. To illustrate this issue, and before we discuss considerations for conducting a dedicated field sampling effort, we'll look at two commonly available forest inventory datasets with differing scales: 1) National scale--the Existing Forest Inventory and Analysis (FIA) plots and, 2) Local scale—Common Stand Exams and Timber Cruises.

National Scale Data: Existing Forest Inventory and Analysis (FIA) plots

Existing Forest Inventory and Analysis (FIA) phase 2 data consist of an established grid of plots, one plot per 6000 acres, for which detailed measurements are made on a 5 year cyclical basis. The measurements are generally made on a cluster of four 1/24th acre subplots (24ft radius) where trees of 5" dbh and larger are measured.

Why FIA plots are insufficient for our purposes:

- Although a consistent field protocol is used to acquire all the FIA plots, the sample density of one sample per 6000 acres provides **too few samples** for the lidar analysis at the project scale.
- There are three **biomass** problems with the FIA protocol when the data are used to establish statistical relationships with their corresponding lidar points:
    1. A field crew applies a set of rules to decide which trees are inside our outside the plot based on the distance of the tree bole to the plot center. Lidar data represent the canopy biomass from above. When an area corresponding to a field plot is subset from the lidar data, all the lidar points within the plot area are included whether the tree bole is inside or outside the plot area. A large tree just outside the plot, can contribute a large amount of biomass to the plot, more so if the tree is leaning across the plot boundary. Field measurements will adjust for this, measurements from the clipped lidar data will not.
    2. The smaller the plot size, the larger the relative **edge effect** is. Experience has shown that the edge effect is too large on 1/24th acre FIA plots. The edge effect becomes acceptably small for 1/10th acre plots or larger.
    3. Another problem with using FIA plots is the minimum tree size of 5 inch dbh that is measured on the micro-plots. Lidar pulses are returned from all biomass in the overstory, not just the larger trees. When trees smaller than 5 inches dbh contribute to the top of the canopy (younger stands or mixed stands, these trees are part of the clipped lidar plot but are not accounted for in the micro-plot FIA inventory.
- FIA measurements are made on a fixed timing schedule which might not corresponds with the lidar acquisitions—this can lead to significant **time discrepancies** because of disturbances such as fire, tree mortality, or silvicultural treatments.
- The locations of FIA plots are not publically known and are generally not measured with sub-meter accuracy.   Sub-meter locations are best to create a good fitting relationship between the two data sets. **Errors in location**, just as a timing discrepancy, can result in attempting to relate two different conditions.

Local scale data: Common Stand Exams and Timber Cruises:

At the local (or project) level, the forest information is typically obtained at the stand level (not plot level). Although many more samples will be available, similar issues as with the FIA plots will be encountered. At the project level, field plots are measured as part of a timber cruise or a stand exam each having a variety of field protocols—thus, not all plots contain the same measurements. This can lead to inconsistencies when plots measured for different purposes are combined over an area.

Why Common Stand Exams and Timber Cruises are insufficient for our purposes:

- Plots focusing on volume are using a **variable radius type plot**. This type of a plot cannot be matched with its equivalent lidar plot because the radius is different for each plot location and the radius is not known.
- The type of information collected and the precision of information depends on the examination level (quick plot, extensive or intensive examination)—thus, the data can lack **consistency**.
- CSE plots are **summarized up to the stand level** not to the plot level. This makes it impossible to use these plots to estimate forest attributes on a per acre basis.
- Just as with the FIA plots, the **time of measurement** and **approximate location** will cause issues when relating the field data to the corresponding plots.
- Timber cruises are done to get a reliable estimate for timber appraisals. The focus is on tallying **commercial timber volume** – not total timber volume nor the total biomass that lidar will estimate.
- Stands exams are usually performed on selected stands based on some management criteria. Not all the conditions present in the project landscape are represented in the stand exams. To be able to develop a relationship between the plot data and the lidar data the **entire range** of conditions must be represented.

As illustrated above, both national-scale and local-scale field data have significant problems correlating with metrics derived from the lidar point clouds. The problems become even more pronounced when combining plots from different inventories. Modeling input data derived from several field data sources will most likely result in a table or database with many No Data records. During the regression modeling process, all records are required to have data for all fields included in the analysis. A table with many sparsely populated records may have taken a long time to compile but will not yield meaningful results. Thus, the effort of collecting dedicated field data remains largely unavoidable to successfully quantify forest attributes from lidar data.

Dedicated Field Sampling For Lidar Derived Forest Inventory Estimation

From the above discussion, it should be clear that better predictions can be obtained with field plots that are measured with the specific goal of relating the lidar and inventory data. Doing so will optimize the information that can be derived from the lidar data. As with any field data collection the question of **how many samples**, **their distribution** and **what to measure** must be determined—i.e. you must have a sampling design and a field protocol.   The objective of your sampling design and field protocol is to discover relationships that will allow you to quantify vegetation structure from lidar pulse height information[2]—while balancing statistical validity with cost restraints. Regression models relate field plot data to lidar plot data and then make predictions across the extent of the lidar data. The accuracy of the predictions is improved if field data are collected across the entire range of variability in the population.

[2] It should be noted that the objective of most sampling efforts is to make some inference about a population (e.g. the mean and standard error of the mean) from the characteristics of the sample. This objective changes when field plots are measured to discover and test the relationships with another data set measuring the same forest attribute(s) by different means (in our case an airborne laser scanner). The goal is no longer to generate estimates of a population but to model and quantify vegetation structure from lidar pulse height information.

We recommend that you consult an experienced statistician or biometrician to develop a sampling design and field protocol—however, a few general guidelines and suggestions follow.

Sample Design

Many sample design options such as random, systematic, stratified, and hybrid are feasible. However, if field locations are determined at random, it is likely that the ends of the data distribution will not be included. The ends of the data distribution have great influence on regression models and should be included for good model fit. In addition, without samples at the ends of the data distribution, model estimates beyond the range of the field collected data are notoriously unreliable. A study by Hawbaker and others, 2009, illustrates that a **stratified sample design**, selected from the lidar data, results in better forest attribute estimates when compared to models based on the analysis of plots selected using a random sampling design. The stratified sample produced a greater range of attribute variability and the predictive regression models minimized extrapolation beyond the range of the observed field data. On the other hand, a stratified sample design may require more sample plots than a random sample design. Again, we recommend that you consult an experienced statistician or biometrician to develop a valid and cost-effective sampling design.

Assuming the lidar data has been acquired and is available, the suggested stratification can be accomplished based on variability of **height**[3] and **canopy cover**[4]. Creating these two raster grids is a fairly straightforward process in FUSION. Both of these metrics are related to how much biomass is present in the forest and some measure of both are frequently showing up as best predictor variables to estimate forest inventory parameters. Existing stand maps, inventory data, vegetation maps and interpreted spectral data (resource photography, NAIP, etc.) can help identify and stratify conditions of interest. The best ancillary information is comparable in scale and should represent similar ground conditions, i.e. as co-temporal to the lidar acquisition as possible.

Plot Protocol

After establishing where and how many field plots to visit, the specifics of what to measure at each plot must be addressed. The objectives and processes will likely be different from what field crews are familiar with. The typical objective of collecting field plot measurements is to make some inference about a population. The objective in this case is to discover and test the relationships with the lidar data. Thus, we need to adjust our field measurements to better reflect how lidar technology samples the field plots. The lidar plot data is essentially a cylinder that includes all of the 3-D lidar returns within a fixed radius of a point location—the field measurements should thus include all biomass features within the same 3-D space even if, for example, the tree stem is outside the plot but the canopy is within.

Not only will adjustments be made to the traditional way of doing field measurements, field crews will have to make adjustments in the field. Specific instructions and training are necessary to effectively handle those occasions when field conditions prevent the field plot data from corresponding to the lidar plot data. Examples include:

- A large tree with the trunk outside the plot radius but the crown taking up a large portion of the plot. Solution: it is better to move the center location of the plot.
- The plot is located in a mixed condition (edge of a burn, open and closed canopy ecotone, etc.). Solution: it is better to move it so the field measurements only represent a single condition.

[3] Several measures of height are available including: mean height, max height, standard deviation, and others.
[4] Canopy cover is often referred to as vegetation density in the lidar community.

Field prep work

Following are suggestions and recommendations to ensure that field time is efficient and cost-effective.

- Create location maps with high spatial resolution imagery (capable of distinguishing individual tree crowns) of the plot sites. ArcMap has a MapBook utility that is very useful to create these maps.
- Create your field sheets (or data loggers loaded with required software).
- Create list of tree species alpha codes (these will be migrated to FVS_spp codes later).
- Decide the units used for the measurement (English or metric) and use the equivalent tools (tapes, calipers,…) – think about the units used in equations or models to derive the parameter of interest.
- Explain to the field crew why the procedures might be different from how things are done for other field studies. Consider showing field crews some lidar plot subsets to help them visually understand the potential issues.

Field protocol for lidar correlated plots

The cost for collecting field data ranges between $500 and $1000 per plot. It adds significant cost to a lidar project. If it is done well, the field data will result in good forest estimations from the lidar data. If it is not done well, the data cannot be used to relate field conditions to the lidar data and the effort will not contribute to the objective of the project.

The lidar scanner samples everything in its view path. In forested areas the returns will represent the vertical forest structure. Under dense canopies, most returns will come from the upper canopy. Thus, the biomass in the upper canopy is more influential when modeling forest structure than the lower vegetation. In fact, returns below two meters are usually filtered from the lidar data before the modeling process starts. These and other issues should be reflected in the field measurement protocol. Following is a list of suggestions that should be considered to make the match between the field measurements and the lidar data as close as possible:

- Field data should best be acquired within 1 growing season of the lidar acquisition.
- Use a fixed radius plot (a variable radius one, although fast to get field measurements of BAF, cannot be clipped from the lidar data since the radius is not known).
- Plot size: minimum 0.1 acres when there are 8 plus trees with at least 3 inch diameter. When there are less than 8 trees with 3+" DBH increase the plot size to 0.2 acres. Edge trees on plots smaller than 0.1 acres can cause difficulty when relating the plots to the lidar data.
- Minimum tree diameter: (3+") – smaller diameter trees can be excluded from the analysis later. Trees less than 3" when representing understory are less likely to be seen in the lidar point cloud.
- Minimum number of trees to measure: 8+ (if there are less but large diameter trees on the plot it is best to expand the plot size or change to location according to protocol instructions. If there are fewer than eight and all small diameter trees, the plot most likely does not represent forested conditions–the conditions may have changed since the plot location was selected).
- Measure dbh and record species code of all the trees larger than the diameter cut-off including the non commercial species – since they do contribute to the biomass on the plot.
- Measure total tree height to the top of the tree– not merchantable height.
- Smaller trees can be counted by size class and species on a smaller subplot coinciding with the center (so the location coordinates are known and can be subset from the lidar data. Smaller trees do not contribute as much biomass reflections to the total point cloud, especially when overgrown by older growth trees.
- For each plotID record:

    - Project name, name of the field tech, and plot ID,
    - plot size, (used to calculate per acre inventory equivalents)
    - time and date of measurement (a backup check to relate GPS locations afterwards),
    - general condition of the plot (poor – soils depleted of all nutrients because of fire might result in stumped trees, rich—favorable water conditions might result in taller trees than average conditions). This description will help when plots don't fit the trend of the regression analysis.
    - Tree list for each plotID:
        - treeID,
        - species (2 letter alpha-code used in FVS),
        - diameter,
        - height for 2 dominant and 2 co-dominant trees (to check against calculated heights or to evaluate if existing regional height diameter equations do not work well for local conditions),
        - condition class (i.e. live/dead),
        - Live Crown Ratio (LCR),
        - Crown class (dominant, co-dominant, intermediate, overtopped or remnant).
- Other vegetation: ocular estimate of cover (example certain percent of plot is shrub while also listing dominant and subdominant species and include an overall height estimate.
- Mark the plot center (just in case it has to be revisited).
- Mark the trees as they are measured with chalk or spray paint (to avoid measuring trees more than once or missing a tree).
- Give field crew instructions related to moving the plot location within reason (to capture a single condition, to take care of large edge trees) and how and when to increase the plot size. A note of caution: plot conditions at a more accessible location might look the same as the conditions of the preselected location. However, the field crew may not be aware of the characteristics that contributed to the plot being selected and substituting the plot conditions of a selected plot with one of their choice can interfere with the goal of capturing the range of variability.
- Make the field crew aware of the benefit of taking good notes about plot and surrounding conditions.
- Take photos at the plots (might help explain problems establishing the relationships between the field data and the lidar variables).

GPS procedures

The three components of ensuring that the field data will correspond to the lidar data are: 1) ensuring the field data is collected within one year of lidar acquisition, 2) ensure that field measurements correspond to the lidar 3-D plot cylinder, and 3) ensure that the field plot and the lidar plot are precisely co-located. This third component--recording accurate plot locations is a very important because if the field data cannot be spatially correlated to their corresponding lidar data metrics, all is lost. Field plot locations should be measured with GPS within 1 meter of their true location.

GPS receivers can be classified into three major groups each with different technological capabilities and corresponding price ranges:

1. <u>Recreational grade receivers</u>. These are inexpensive receivers but also the least accurate. They are unacceptable for correlating the ground plot locations with the lidar plot locations. They can be used to navigate to the general vicinity of the plot—but should not be used for anything more.

2. Mapping grade receivers. While not as accurate as the survey grade receivers (discussed next), mapping grade receivers strike a good balance between accuracy, ease-of-use, and cost. Typically, they are the receiver of choice for this type of project. Using standard GPS data collection procedures, all mapping grade receivers will typically produce positions with under-canopy accuracies that are well within 5 meters of the true location. If the mapping grade receiver is capable of receiving and processing both L1 and L2 signals (this requires an additional external L1/L2 antenna), the under-canopy accuracy should be within a meter of the true location—this is the recommended accuracy and recommended GPS configuration. Note: even with the dual frequency (L1/L2) receivers, you must differentially correct the GPS data to achieve the desired accuracy.
3. Survey grade receivers. This type of receiver exceeds the accuracy required for this type of project. Given the high cost, steep learning curve, and more stringent use requirements, survey grade receivers are not considered to be the best choice.

Assistance is available for current hardware recommendations, training, and support through the Field Data Automation—Mobile Computing (FDAMC) website:

http://fsweb.wo.fs.fed.us/irm/fdamc/

In addition, GPS and Mobile GIS help is available through the Forest Service Helpdesk.

Plot Data processing – preparing the forestry and corresponding lidar variables for modeling

After all the data are collected, they need to be processed and prepared for modeling. The goal is to process the field inventory and the lidar data to ensure they correspond as much as possible. This will entail summarizing the field inventory data to the plot level (e.g. instead of a height measurement for each tree in the plot, a single height value for the entire plot will be calculated), and creating corresponding lidar metrics from the plot locations. The end product of this step of the workflow will be a single flat table that contains a record for each of the field plots. Each plot record will include the field plot information, the GPS location, and the corresponding lidar metrics. The number of fields or columns depends on which attributes will be predicted (Table 1).

The final flat table needs to have every field (cell, variable) populated for every record (plotID, or row). Blank cells or nodata cells are not allowed for either the predictor or the response variables during the linear regression modeling – a single blank cell for a record will result in that record being excluded from the analysis. The time required to clean and fully populate the final table is usually underestimated.

Table 1: Type of variables in the flat table that will become the model input table:

Data Source	**Corresponding fields**
1. Field plots	Plot ID, forest variables to be estimated (inventory variables, biomass, fuels, other).
2. GPS data	XY locations. These are not required for modeling; they are only used to facilitate linking the field measurements to the lidar data. A significant advantage of keeping the XY locations in the table is that it allows you to spatially display the data in ArcMap.

3.Lidar plot data	Lidar cloud metrics divided in the following categories: Return counts above the canopy cut-off (depends on max number returns per pulse recorded) statistical descriptors (16 fields), canopy density enumerators (12 fields). Note these data are based on version 2.90 of FUSION (McGaughey 2010).

The remainder of this section provides brief descriptions of the post-processing methodology required for each of the 3 data sets. Preparation of the field plot data and GPS data can be accomplished in a variety of software packages. Generating the lidar plot data from the lidar data will be described using the FUSION software package.

Generating forest modeling variables from the field plot data

To be able to use the field measurements, it is required to convert the measurements for each tree to the desired attributes. If the data were captured in a table, spreadsheet of inventory specific program on PDA or a data logger, it is only as matter of transferring the data and formatting the already digital information into the right format compatible with the processing software that will be used (FSVeg, FVS, etc.).

The following information is required for each tree at every ground plot:

- Plot number
- Tree ID
- Plot size
- 2-digit tree species code
- DBH
- Live/death (coded by live = 1 and death = 0)

In preparation for modeling, the measurements for each tree (above) must be converted to the desired tree attributes (described below) and then summarized to the plot level. This can be a complex process for the uninitiated. We recommend that you recruit local expertise if your team does not already possess these skills.

When local expertise is not available, a spreadsheet can be created that can do the job--and the learning curve for the spreadsheet will not be as steep as learning an inventory-centric software package. This procedure requires downloading an Excel DLL plug-in from the National Volume Estimator Library (NVEL), and installing it (http://www.fs.fed.us/fmsc/measure/volume/nvel/index.php). For each species measured in the field, the equation that is most suited to calculate the tree volume needs to be extracted from the database.

Since the calculated volumes depend on the species specific tree height, the height needs to be calculated as well. The heights are calculated from the DBH measurements according to a height-diameter formula. These formulas and the required coefficients are species and region specific. The coefficients can be found in section 4.1 of one of the 20 regional FVS variant documents available at: http://www.fs.fed.us/fmsc/fvs/variants/index.shtml

Desired Tree Attributes. The following tree attributes will be calculated and/or used based on the ground plot tree measurements:

- Tree height(ft)
- Basal Area (sqft/ac)
- Live BA
- TPA and live TPA
- Ht * Live_BA (used in the next plot summary section to calculate Lorey's height)

- Live tree volume data
    - live volume in cubic feet (cuft) per tree
    - live volume in merchantable board feet (bdft) per tree

The ground plot tree measurements have to be summarized to the plot level by operations in Table 2.

Table 2: 'Tree fields' needed to summarize to the plot level

Input field	Summary operation
Plot-ID	known
Plot size	known
BA	sum
Live BA	sum
TPA (trees/acre)	sum
Live TPA	sum
Ht * Live BA	sum
Live volume (cuft)/tree	sum
Live volume (bdft/tree)	sum

Finally, you may want to calculate four more variables in the plot level data sheet[5]:

Variable	Formula
Lorey's height	= (Ht * live_BA) / Live_BA-cell
QMD	=(((Live_BA_sqft_ac / Live_TPA)/PI())^0.5)*2*12
Live volume cuft per acre	=(Live_Vol_cuft_tree /plotsize)
Live merchantable volume bdft per acre	=(Live_Merch_bdft_tree/plotsize)
Note: QMD ^0.5 = square root; 2*12 are conversion factors from radius to diameter (2) and feet to inches (12)	

This concludes calculating the forest inventory variables from the field measurements and summarizing them to the plot level. The forest variables that will be estimated at the landscape scale need to be exported to a separate flat table, including the plot-ID. They will be joined with the corresponding plot lidar derived metrics at the end of this section.

Post processing GPS data

The GPS data must be differentially corrected. Note: since these are point data, the average of all the points measured at each of the plot locations will yield the most accurate position. The positions need to be converted to the same datum, units, and planimetric coordinate system as the lidar data.

The last step involves checking the plot ID (or assigning it, if that was not done before) and confirming the date and time for the GPS location corresponds to the time and date the plot was actually measured in the field (by checking the field notes). Linking all the pieces of information together is best done using a relational table link in a database (Access or ArcGIS). The plot-ID will become the key field used to link the pieces of information (forest variables, XY GPS locations and lidar metrics) together. Making sure the plot-ID is named exactly the same way in the 3 datasets will make it easier to relate the data into a single record.

[5] The four additional variables documented are simply examples, many more could be potentially generated.

Generating lidar predictor variables

At the plot scale

Once the XY coordinates for each of the plots are available, the last step in the data preparation process consists of subsetting the lidar returns that correspond to each field plot. During the subsetting process the data are normalized to the ground surface so the returns are expressed in terms of heights above the ground instead of in elevation. After subsetting the lidar plot equivalents, the last step consists of calculating a set of 47 cloud metrics variables for each of the plots. All but one is calculated using the FUSION function, the remaining one is derived in a spreadsheet. These variables will be used as the predictor variables in the linear regression modeling.

Clipping the area corresponding to each field plot from the lidar data

This step in the workflow is fairly straightforward once the single batch file that subsets all the ground plots is created. Creating this batch file might look daunting at first glance but is not that hard once all the needed components listed below are in place:

1. A list of the plot-IDs with their lower left and upper right bounding box coordinates
2. A list of all the high spatial resolution bare earth surfaces, usually delivered by the vendor
3. A list of all the LAS files

The batch file will have a separate command line for each of the plots that is clipped from the lidar acquisition. Each line in the batch file has to following basic structure (McGaughey 2010):

```
ClipData [switches] InputSpecifier SampleFile [MinX MinY MaxX MaxY]
```

A more detailed workflow is described in the exercises.

Generating cloud metrics for the lidar plots

Once the lidar returns corresponding to the ground plots are extracted, the last step includes summarizing the lidar returns for each plot in a set of variables representative of the vertical distribution of the forest structure. These numerically summarizing variables make it possible to describe lidar plots during the analytical and quantitative modeling. Details of the FUSION syntax are listed in the Fusion Manual. The command has the following structure (McGaughey 2010):

```
CloudMetrics [switches] InputDataSpecifier OutputFileName
```

Each record in the output cloudmetrics.csv has a set of variables (fields) that together describe the vertical distribution of the lidar points (representative of the biomass) within the plot (Table 3). Once the cloudmetrics file is generated, the lidar data preparation is finished.

Table 3: Groups of lidar plot variables generated by cloudmetrics (McGaughey 2010):

Category	Output variable
Descriptive	Total number of returns Count of returns by return number Minimum Maximum Mean Median (output as 50th percentile) Mode Standard deviation Variance Coefficient of variation Interquartile distance Skewness

	Kurtosis AAD (Average Absolute Deviation) L-moments (L1, L2, L3, L4) L-moment skewness L-moment kurtosis
Height percentile values	(1st, 5th, 10th , 20th, 25th, 30th, 40th, 50th, 60th, 70th, 75th, 80th, 90th, 95th, 99th percentiles)
Canopy related metrics (calculated when the /above:# switch is used	Percentage of first returns above a specified height (canopy cover estimate) Percentage of first returns above the mean height/elevation Percentage of first returns above the mode height/elevation Percentage of all returns above a specified height Percentage of all returns above the mean height/elevation Percentage of all returns above the mode height/elevation Number of returns above a specified height / total first returns * 100 Number of returns above the mean height / total first returns * 100 Number of returns above the mode height / total first returns * 100
Others	See FUSION manual

Generating the table with the to-be-estimated forest variables and the lidar predictor variables

Before generating the predictive models from the data described in this section, one last step is required: combining both the forest variables derived from the field data and the cloudmetrics variables from their corresponding lidar plot into a single flat table. This table will have as many records as there were usable field plots. Every cell in this table should have a value. The XY locations are not required for the modeling, however, the table can be used in ArcGIS if they are included. The best way to join the data into a single table is by relating the different pieces (field data and lidar data) based on the plot-ID. It can be done in a spreadsheet but there is significant potential for making mistakes. When a spreadsheet is used, both tables should be ordered in the same order using the plot-id (make sure the same naming convention is used in both tables and that there is no ambiguity in the plot names). Both sets of variables can be pasted into the table and the final output table is ready to start the modeling.

Generate Lidar Metrics for the Landscape

This step in the process can be done later if desired; however, we discuss it here to draw your attention to the relationship of Fusion's cloudmetrics command and Fusion's gridmetrics command. While there are significant differences in how the two Fusion commands are implemented, they generate the same output variables. Cloudmetrics generates its output from the lidar cloud within the boundaries of the 3-D plot while Gridmetrics generates the same output based on the lidar cloud within the boundaries of each 3-D grid cell—across an entire grid. In other words, the output of Gridmetrics is the same as Cloudmetrics but it will be a continuous raster grid for each of the output variables.

After the regression models are developed, these grids will be the input variables or the predictor variables, to which the models are applied resulting in an estimated forest attribute of interest - also in grid format - at the landscape scale.

One of the Gridmetrics parameters is the cell size of the final raster data. Currently it is recommended to select a cell size that corresponds to the area of the field plots (because that was the spatial area used to establish the relationship and has been shown to work)—Table 4. It might be possible to extrapolate the relationships to another cell size but research to prove this works or how much error this might introduce are not yet available.

Table 4: relationship of 0.1 and 0.2 acre plots to gridmetrics cell size

Plot size	Radius in ft.	Cell size	Radius in m	Cell size
0.1 acres (1/10th) (1 acres = 10 chains)	37.23 ft	1 chain^2 (1 chain=66ft.)	11.34 m	20 m
0.2 acres (1/5th) *only used when there are less than 8 dominant and co-dominant trees on the plot.*	52/66 ft.	NA	16/05 m	NA

The layers generated by this process are the same as those from the cloudmetrics. When topographic layers are used, additional layers can be created. For additional details and specifics of which columns to extract out of the intermediary CSV files check the fusion manual (McGaughey 2010) and available RSAC training modules.

Data processing conclusion

This section illustrated how the plot data, GPS data, and the lidar data should be processed at the plot scale--and the landscape scale for the lidar data. The plot level data are used to develop the models to estimate the forest inventory variables while lidar metrics at the landscape scale will be used to apply the regression models at the landscape scale.

Developing Statistical Models.

Developing a valid statistical model is a complex task and detailed instructions for developing statistical models are far beyond the scope of this document. We will, however, provide general considerations and guidelines that will help ensure a statistical model that represents conditions in the field. Our first suggestion: if statistical expertise is lacking within your team, we strongly recommend that you consult a statistician or biometrician for this portion of your project.

While there are a number of options for developing statistical models, we'll focus primarily on linear regression[6]. Detailed regression equations with their summary tables and corresponding graphs for forestry inventory models is documented in the Colville National Forest project report (Reutebuch and others 2010). In addition, the use of regression equations for biomass and fuels estimations is documented by Andersen and others (2005).

Existing studies indicate linear regression works best in conifer dominated forests; the predictions are not as good in mixed forests. The models cannot be applied universally and have to be developed on an area by area basis.

Linear regression is a parametric statistical method. Parametric methods provide superior models to non-parametric methods—but, parametric methods make several important assumptions about the underlying data (normality, homogeneity of variance, independence, etc.). Exploratory data analysis and

[6] All of the analysis can be accomplished in the R environment (http://cran.r-project.org/). R is an open source statistical software package which becomes a lot easier to use with a few additional open source add-ons such as TINN-R (data editor), R-commander (GUI for many statistical functions) and Rattle (GUI for data mining). In addition, the Remote Sensing Applications Center has developed lidar data-analysis tools to streamline the process of building statistical models. The tools guide the analyst through model building and provide tests of the statistical validity of relationships between field plot and lidar data. These tools include options for using regression modeling or Random Forests to model forest inventory parameters thus, the tool is flexible while also offering safeguards to avoid using inappropriate statistical-analysis techniques.

computation of univariate statistics should be conducted prior to linear regression analysis to see if parametric assumptions are valid and to get a general feel for the data. If you should discover that the parametric assumptions are violated, there are mathematical transformations that may be used to force the response variables into a normal distribution.

The objective of modeling is to define the best equation that represents the trend between the two sets of variables and represents reality. Care should be taken to fit the trend and not over-fit the individual data points. During the regression analysis using the R-squared value is a good gage for the model fit but the quest for the best R-squared can lead to over-fitting the data. The best model follows the general principles of parsimony:

- Models should have as few parameters as possible,
- Linear models should be preferred to non-linear models,
- Experiments relying on few assumptions should be preferred to those relying on many,
- Models should be pared down until they are minimal but adequate,
- Simple explanations should be preferred to complex explanations.

Most predictive lidar based models should not have more than three variables generally representing some form of the three metrics listed below:

- One related to height (a percentile variable),
- One related to canopy cover and,
- One describing the variation in the data (standard deviation or variance).

Another important requirement for an appropriate linear regression model is that the data are related linearly. If that assumption is violated, there may again be transformations that can be applied to create a more linear relationship. Even when the data are linearly related, there is danger in extrapolating the modeled relationship beyond the range of the field-collected data. Extrapolation problems are addressed by following the critical characteristics of the forest inventory data (outlined previously): you must have enough plots for statistical validity and the plots must cover the full range of variability of the measurement of interest.

Linear regression modeling – generalized workflow in R

The following table outlines the regression model workflow (Table 5). The workflow is similar when other software packages are used but the specific R-commands listed in the table will obviously not work.

Table 5. Regression Model Workflow for R.

Step #	Procedure or Command	Explanation
1	Run a Best subset regression (BSR) model for all the forest response variables	Check to see which are common predictor variables to models or where substitutions can be made. This reduces the number of predictor variables to work with
2	Pick the first forest variable and run a BSR using the subset of the predictor variables	Pick the best linear models from the output table
3	Run the linear regression model selected in the step above Model<-lm(response~predicted variables)	

4	Evaluate the model output Summary(model)	

5	Plot the diagnostic graphs
	oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
	plot(model)
	par(oldpar)
		Check visually for regression assumptions and outliers and unusual data points

6	If there are outliers (visual and Bonferroni test), rerun the model excluding the observations deemed outliers -c(observation), next run –c(observation1, observation2)…	Check for regression assumptions and outliers
7	If the Q-Q-plot of the residuals is not normally distributed, check if a boxcox transformation would resolve this Boxcox(model)	Evaluate the graph and determine the power of the transform
8	Transform the response variable, rerun the lm model Transformed Model<-(transformed-y~ predictor variables), create summary and plots	Evaluate if the errors are normal and there are no more outliers
9	If there is more than one predictor variable, check the variance inflation factor Vif(transformed Model)	Make sure there is no collinearity between the predictor variables
10	Once an acceptable model is generated, extract the information required to build the regression equation from the model summary table as well as the R-squared and the adj R-squared (to plot onto the graph in step 13)	Build the regression equation
11	Extract the model RSE	Backtransform the regression equation and calculate the correction factor
12	Start over from step 2 and repeat the procedure for the next forest variable. If more outliers are detected while developing subsequent models, the previous model(s) need to be regenerated also excluding the outliers from following models.	

13	Once all the desired models from the same data set are generated and all the regression equations are built, there is one more step to do: plot the observed values against the plotted values.
	
	response_Pred<-fitted(model)
	response_Obs<-(transform(file$column)[-c(outlier_observations)])
	plot(lmodel_Obs,model_Pred)
	lines(c(0,100000), c(0,100000))
	title(main="transform(model)~predictor variables", sub="Residual standard error: ####; Adjusted R-squared:
		Creates a scatterplot of observed values against the fitted values. This shows visually the fit of the model.

	####")	

Summary—Developing the Statistical Model

At the end of the regression analysis, there will be one predictive model for each of the forest variables that were derived from the field measurements. The fit of the models should be high (in the 0.6 to 0.8 adjusted R-squared range) if the field data and the location data were measured accurately and the models were developed employing good forestry principles in combination with good statistical judgment[7].

A Final Note: RandomForest classification provides an alternative to linear regression modeling. While it still requires statistical expertise, the process can be far less complex—especially if the RSAC data analysis tools are used. However, this approach has been used far less than linear regression modeling and the cost-benefits are not fully known yet.

Generate Estimated Forest Inventory Data at the Landscape Scale

At this point in the process all of the heavy lifting has been completed—you're almost ready to apply the models that you've created. However, prior to applying the models, you should use the data you have to create a Forest/non-Forest Mask. The mask is desirable since areas without trees are of little interest to estimate forest attributes and the mask will reduce errors. In young forests characterized by recent regeneration the predictions are usually below the range of field measured data. Under these conditions estimated attributes are extrapolated outside the range of values from which the models were created and the results can be highly erroneous. Non-forested areas and areas with immature forests are better excluded from the inventory predictions by creating a Forest/Non-forest mask.

The forest/non-forest mask layer is obtained by combining the results of 2 separate conditions:

Condition	**Remark**
Canopy Cover ≥ 2 %	This value can be set to any value that is appropriate for local conditions
90th Elevation percentile height	often canopy cut-off value used for the project ≥ 3m or 10ft

Some vegetation mapping applications use a forest/non-forest cut-off value of 10% for canopy cover. A stricter cut-off value can always be imposed on the estimated attributes after they are generated. Once each of the individual forest masks is created, it is important that cells with zero values are reclassified to 'NoData' and each layer is saved or exported with a meaningful and recognizable grid name. A combined forest/non-forest mask is then easily created by multiplying the grids of the 2 previous conditions. The last preparatory step consists of visually checking the mask against another source of high resolution imagery such as NAIP (ideally acquired during the same year as the year the lidar data were collected) making sure that areas for which no forest attributes will be estimated are indeed non forested areas. If this visual check is satisfactory, the final mask should be saved and the spatial analysis mask should be set to this forest/non-forest mask.

[7] Even if the best procedures are used, there remains the possibility that the lidar metrics are not closely related to the corresponding field measurements and the modeling efforts will not be successful.

Estimating Forest Attributes At the Landscape Level: Applying the Models in ArcGIS

Applying the statistical models in ArcGIS is likely the fastest step to perform of the entire analysis process. The statistically-derived equations from the modeling process are used in the ArcGIS Spatial Analyst environment. However, instead of the plot cloud metrics predictor variables that were used to generate the equations, you will substitute the corresponding grid metric layers. The output of each calculation is a new grid in which each cell spatially represents the estimated variable of interest derived from the lidar data. The output regression equations for six response variables is provided in the following table (Table 6).

Table 6. Example regression equations for six response variables.

	Response variable	**Regression Equations**
1	LHT_ft	= 24.41 + 0.753(ElevP80)
2	(LBA_3in_sqftac)	= sqr(-5.11 + 0.198*(ElevP90) -0.2777*(ElevSD) + 0.114*(PC1stRtsCC))+3.003
3	(LTPA_3in)	= exp(3.67 -0.005*(ElevP80) +0.029*(PC1stRtsCC))*1.108
4	(LQMD)	= exp(1.68 + 0.015*(ElevP80) -0.004*(PC1stRtsCC))*1.033
5	(LV_cuftac)	= sqr(-52.52 +0.954*(ElevP90) + 0.647*(PC1stRtsCC))+150.660
6	(LMV_bfacc)	= sqr9-138.68 +2.452*(ElevP90) + 1.399*(PC1stRtsCC))+952.957

Basic Quality Check Of The Estimated Attributes.

Short of going into the field to check the prediction outputs, the result can be evaluated using a few simple steps:

- Take a few random locations and check the raster values against the corresponding lidar point cloud. Although this is not a direct quantitative comparison, the height and the vertical distribution of the points should be indicative of the feasibility: examples raster cells with high volumetric values should correspond to field conditions that can yield these numbers.
- Compare the data range of each of the predicted variables to their ground plot ranges. The differences can be represented as percent differences in tabular form.
- Pixels having values either 10% less or more outside the plot range can be flagged using a conditional statement. These grids can be incorporated in further models or analysis.

Deriving Second Generation Forest Attribute Layers

Once the models have been successfully run, the results can be used as input variables for other models or to calculate spatially explicit forest information that uses these variables as input. The principle is the same as the one used to predict the forest attributes listed previously, the only difference being that the estimated attribute grids are substituted in the raster calculator equations for lidar metric grids as shown in the examples listed in the following table (Table 7).

Table 7: 2nd generation derived forest attributes based on the initial set of lidar inventory forest ones

Index	Formula
SCD (stand density index)	TPA(QMD/10)^1.604
QMD (based on predicted BA)	sqrt(BA/(0.005454*TPA) when BA is in sqft/ac and qmd in inches
RD (Curtis' Relative Density)	BA/sqrt(QMD)

All these data sets can be used for further ecological, habitat and other resources related modeling.

References

Andersen, Hans-Erik; Robert J. McGaughey; Stephen E. Reutebuch, 2005. Estimating forest canopy fuel parameters using LIDAR data. Remote Sensing of the Environment. Vol.94:p 441–449

Andersen, Hans-Erik; Clarkin Tobey; Winterberger Ken; and Strunk Jacob. 2009 An accuracy assessment of positions obtained using survey- and recreation-grade global positioning system receivers across a range of forest condition within the Tanana valley on interior Alaska. West. J. Appl. For. Vol.24(3): p128-136.

Avery, Thomas Eugene; and Harold E. Burkhart. 2002. Forest measurements 5th ed., McGraw-Hill. 456 pages.

Bechtold, William A.; Patterson, Paul L.; Editors. 2005. The enhanced forest inventory and analysis program - national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p. [http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.pdf].

Bolstad, P.; A. Jenks; J. Berkin; K. Horne; and W.H. Reasings. 2005. A comparison of autonomous, WAAS, real-time, and post-processed global positioning systems (GPS) accuracies in northern forests. North. J. Appl. For. 22(1):p5–11.

Box, G.E.P., and Cox, D.R. 1964. An Analysis of Transformations. Journal of the Royal Statistical Society, B, Vol. 26: p 211-246.

Crawley Michael J. 2009. The R book. Wiley. 942p.

Deckert, C.J.; and P.V. Bolstad. 1996. Forest canopy, terrain, and distance effects on global positioning system point accuracy. PERS Vol. 62:317–321.

Draper, Norman R.; and Smith Harry. 1998. Applied Regression Analysis. 3rdedition. Wiley New York. 706p.

El-Rabbani, Ahmed. 2006. Introduction to GPS - The Global Positioning System. 2nd ed. Artech House, Boston. 210 p.

Fox John, 2002. An R and S-plus companion to applied regression. Sage publications. 311 pages.

Freese, Frank. 1962. Elementary Forest Sampling Agriculture Handbook No. 232. U.S. Department of Agriculture, USDA Forest Service. 91 pages. http://www.fs.fed.us/fmsc/ftp/measure/cruising/other/docs/AgHbk232.pdf

Gatziolis, Demetrios; Andersen, Hans-Erik. 2008. A guide to LIDAR data acquisition and processing for the forests of the Pacific Northwest. . Gen. Tech. Rep. PNW-GTR-768. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 32 p. [http://www.treesearch.fs.fed.us/pubs/30652]

Hawbaker, T. J.; N. S. Keuler; A. A. Lesak; T. Gobakken; K. Contrucci; and V. C. Radeloff (2009), Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design, J. Geophys. Res., 114, G00E04, doi:10.1029/2008JG000870.
[http://silvis.forest.wisc.edu/Publications/PDFs/Hawbaker_etal_2009_JGR.pdf]

Hawbaker, T.J.; T. Gobakken; A. Lesak; E. Trømborg; K. Contrucci; and V.C. Radeloff. 2009. LIDAR-based forest inventory of uneven-aged mixed hardwood forests. Forest Science*, V*ol.53(3): p 313-326.

Hudak, Andrew.T.; Crookston N.L.; Evans J.S.; Falkowski M.J.; Smith A.M.S.; Gessler P.E.; and Morgan P. 2006. Regression modeling and mapping coniferous forest basal area and tree density from discrete-retrunlidar and multispectral satellite data. Can. J. Remoste Sensing. Vol.32(2):p 126-138.

Jenkins, J.C.; Chojnacky, D.C.; Heath, L.S.; Birdsey, R.A. 2003. National-scale biomass estimation for United States tree species. Forest Science. Vol. 49(1): p12-35.

Johnson, Chris and Barton, Christopher. 2004. Where in the world are my field plots? Using GPS effectively in environmental field studies. Frontiers in Ecology and the Environment; Vol. 2(9): p 475-482

Miller Don M. 1984. Reducing transformation bias in curve fitting. The American Statistician. Vol28(2):p 124-126.

Naesset, Erik. 2001 Effects of differential Single- and dual-Frequency GPS and GLONASS observations on point Accuracy under forest canopies. PERS Vol.67(9): p1021-1026.

Naesset, E.; Gobakken, T.; Holmgren, J.; Hyyppa, J.; Hyyppa, J.; Maltamo, M.; Nilsson, M.; Olsson, H.; Persson, A.; Doderman, U. 2004. Laser scanning of forest resources: the Nordic experience. Scandinavian Journal of Forest Research. Vol. 19: p482–499.

McGaughey R.J. 2010. Forest Inventory Modeling Using LIDAR Canopy Metrics. Lidar workshop, April 27, 2010, Salt Lake City, UT. Powerpoint presentation on file at Pacific Northwest Research Station.

McGaughey, R. 2010. FUSION/LDV: software for lidar data analysis and visualization. Version 2.90. Seattle, WA: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station [online]. Available http:// forsys.cfr.washington.edu/fusion/fusionlatest.html.
[http://forsys.cfr.washington.edu/fusion/FUSION_manual.pdf]

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Reutebuch, S.E.; McGaughey R.J.; and Strunk J.L.; 2010. Sherman Pass LIDAR Forest Inventory Project. United States Department of Agriculture, Forest Service. Pacific Northwest Research Station. 80p.

Schreuder, Hans T.; Ernst, Richard; Ramirez-Maldonado, Hugo. 2004. Statistical techniques for sampling and monitoring natural resources. Gen. Tech. Rep. RMRS-GTR-126. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 111 p.
[http://www.fs.fed.us/rm/pubs/rmrs_gtr126.html]

Shiver, Barry; and Borders, Bruce.1996. Sampling techniques for Forest Resource inventory. Wiley and Sons Inc. New York 356p.

Sprugel D.G. 1983. Correcting for bias in log-transformed allometric equations. Ecology, Vol64(1): p209-210.

Ter-Mikaelian, Michael T.; and Michael D. Korzukhin, 1997. Biomass equations for sixty-five North American tree species. Forest Ecology and Management vol. 97:p I-24

U.S. Department of Agriculture (USDA), Forest Service. 2004a. Common Stand Exam Users Guide, V.1.6. Washington, DC: Department of Agriculture, Forest Service, Natural Resource Information System.

Wing, Michael 2008. Keeping pace with global positioning system technology in the forest. Journal of Forestry 106(6):332-338

Wing, M.G.; A. Eklund; J. Sessions; and R. Karsky. 2008. Horizontal measurement performance of five mapping-grade GPS receiver configurations in several forested settings. Western Journal of Applied Forestry 23(3):166-171.