Understand uncertainty

Learning Objectives

After completing this tutorial, you will be able to:

- Be able to list atleast 3 sources of uncertainty / error associated with remote sensing data.
- Be able to interpret a scatter plot that compares remote sensing values with field measured values to determine how "well" the two metrics compare.
- Be able to describe 1-3 ways to better understand sources of error associated with a comparison between remote sensing values with field measured values.

What you need

You will need a computer with internet access to complete this lesson and the data for week 5 of the course.

Understanding uncertainty and error.

It is important to consider error and uncertainty when presenting scientific results. Most measurements that we make - be they from instruments or humans - have uncertainty associated with them. We will discuss what that means, below.

Uncertainty

Uncertainty: Uncertainty quantifies the range of values within which the value of the measure falls within - within a specified level of confidence. The uncertainty quantitatively indicates the "quality" of your measurement. It answers the question: "how well does the result represent the value of the quantity being measured?"

Tree height measurement example

So for example let's pretend that we measured the height of a tree 10 times. Each time our tree height measurement may be slightly different? Why? Because maybe each time we visually determined the top of the tree to be in a slightly different place. Or maybe there was wind that day during measurements that caused the tree to shift as we measured it yielding a slightly different height each time. or. . . what other reasons can you think of that might impact tree height measurements?

When we measure tree height by hand, many different variables may impact the accuracy and precision of our results. Source: http://www.haddenham.net/newsroom/guess-tree-height.html

What is the true value?

So you may be wondering, what is the true height of our tree? In the cause of a tree in a forest, it's very difficult to determine the true height. So we accept that there will be some variation in our measurements and we measure the tree over and over again until we understand the range of heights that we are likely to get when we measure the tree.

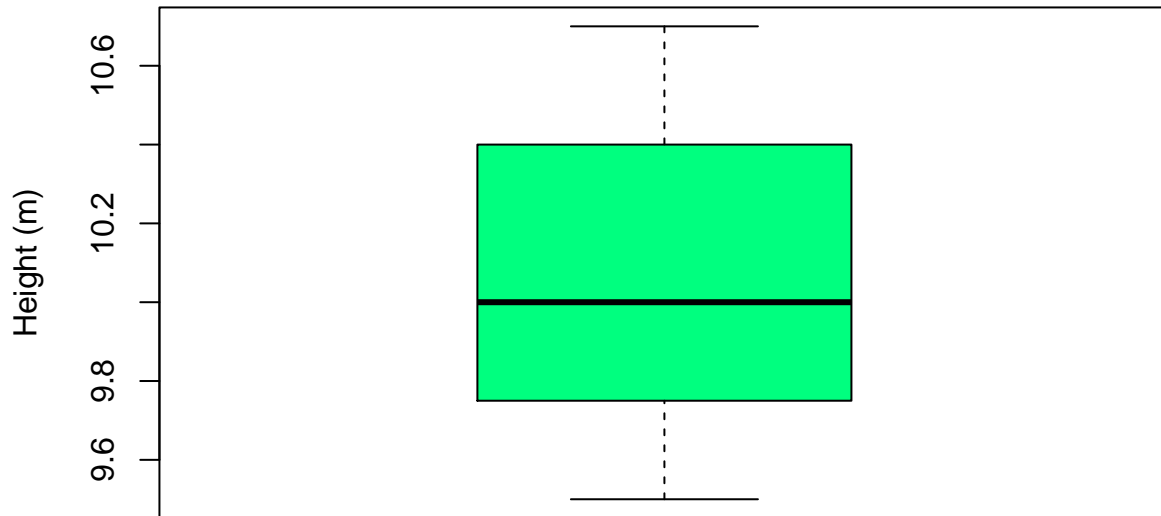**Distribution of tree height measurements (m)**

Figure 1: Distribution of tree heights.

```r
# create data frame containing made up tree heights
tree_heights <- data.frame(heights=c(10, 10.1, 9.9, 9.5, 9.7, 9.8,
                                      9.6, 10.5, 10.7, 10.3, 10.6))
# what is the average tree height
mean(tree_heights$heights)
## [1] 10.06364
# what is the standard deviation of measurements?
sd(tree_heights$heights)
## [1] 0.4129715
boxplot(tree_heights$heights,
        main="Distribution of tree height measurements (m)",
        ylab="Height (m)",
        col="springgreen")
```

In the example above, our mean tree height value is towards the center of our distribution of measured heights. We might expect that the sample mean of our observations provides a reasonable estimate of the true value. The variation among our measured values may also provide some information about the precision (or lack thereof) of the measurement process.

Read more about the basics of a box plot

```r
# view distribution of tree height values
hist(tree_heights$heights, breaks=c(9,9.6,10.4,11),
     main="Distribution of measured tree height values",
     xlab="Height (m)", col="purple")
```

Measurement accuracy

Measurement **accuracy** is a concept that relates to whether there is bias in measurements, i.e. whether the expected value of our observations is close to the true value. For low accuracy measurements, we may collect many observations, and the mean of those observations may not provide a good measure of the truth (e.g., the
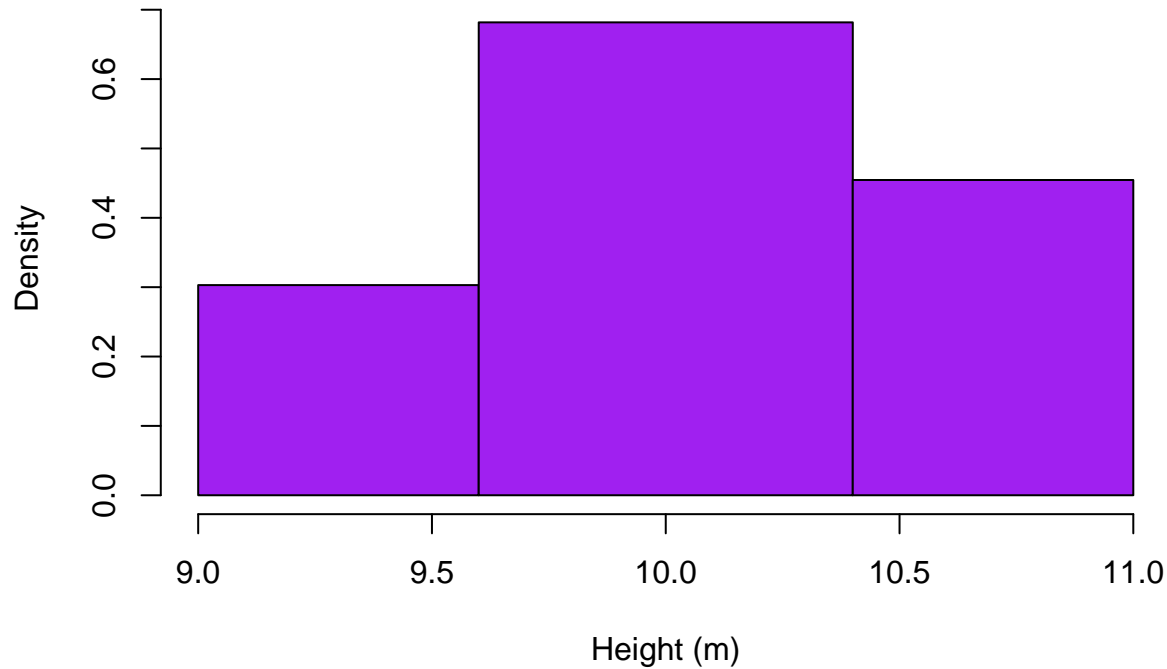
Distribution of measured tree height values

Figure 2: Tree height distribution

height of the tree). For high accuracy measurements, the mean of many observations would provide a good measure of the true value. This is different from **precision**, which typically refers to the variation among observations. Accuracy and precision are not always tightly coupled. It is possible to have measurements that are very precise but inaccurate, very imprecise but accurate, etc.

Systematic vs Random error

Systematic error: a systematic error is one that tends to shift all measurements in a systematic way. This means that the mean value of a set of measurements is consistently displaced or varied in a predictable way, leading to inaccurate observations. Causes of systematic errors may be known or unknown but should always be corrected for when present. For instance, no instrument can ever be calibrated perfectly, so when a group of measurements systematically differ from the value of a standard reference specimen, an adjustment in the values should be made. Systematic error can be corrected for only when the "true value" (such as the value assigned to a calibration or reference specimen) is known.

Example: Remote sensing instruments need to be calibrated. For example a laser in a lidar system may be tested in a lab to ensure that the distribution of output light energy is consistent every time the laser "fires".

Random error: is a component of the total error which, in the course of a number of measurements, varies in an unpredictable way. It is not possible to correct for random error. Random errors can occur for a variety of reasons such as:

- Lack of equipment sensitivity. An instrument may not be able to respond to or indicate a change in some quantity that is too small or the observer may not be able to discern the change.
- Noise in the measurement. Noise is extraneous disturbances that are unpredictable or random and cannot be completely accounted for.

- Imprecise definition. It is difficult to exactly define the dimensions of a object. For example, it is difficult to determine the ends of a crack with measuring its length. Two people may likely pick two different starting and ending points.

Example: random error may be introduced when we measure tree heights as discussed above.

- Source: nde-ed.org

Accuracy vs precision. Accuracy quantifies how close a measured value is to the true value. Precision quantifies how close two or more measurements agree with each other (how quantitatively repeatable are the results) Source: http://www.ece.rochester.edu/courses/ECE111/error_uncertainty.pdf

Using lidar to estimate tree height

We use lidar data to estimate tree height because it is an efficient way to measure large areas of trees (forests) quantitatively. However, we can process the lidar data in many different ways to estimate height. Which method most closely represents the actual heights of the trees on the ground?

It can be difficult to measure the true height of trees! Often times "seeing" the very top of the tree where it is tallest is not possible from the ground - especially in dense, tall forests. One can imagine the amount of uncertainty that is thus introduced when we try to estimate the true height of trees! Image Source: National Geographic

An example LiDAR waveform. Source: NEON, Boulder, CO.

Cross section showing LiDAR point cloud data (above) and the corresponding landscape profile (below). Graphic: Leah A. Wasser

Study site location

To answer the question above, let's look at some data from a study site location in California - the San Joaquin Experimental range field site. You can see the field site location on the map below.

Study area plots

At this study site, we have both lidar data - specifically a canopy height model that was processed by NEON (National Ecological Observatory Network). We also have some "ground truth" data. That is we have measured tree height values collected at a set of field site plots by technicians at NEON. We will call these measured values *in situ* measurements.

A map of our study plots is below overlaid on top of the canopy height mode.

Compare lidar derived height to in situ measurements

We can compare maximum tree height values at each plot to the maximum pixel value in our CHM for each plot. To do this, we define the geographic boundary of our plot using a polygon - in the case below we use a circle as the boundary. We then extract the raster cell values for each circle and calculate the max value for all of the pixels that fall within the plot area.

Then, we calculate the max height of our measured plot tree height data.

Finally we compare the two using a scatter plot to see how closely the data relate. Do they follow a 1:1 line? Do the data diverge from a 1:1 relationship?
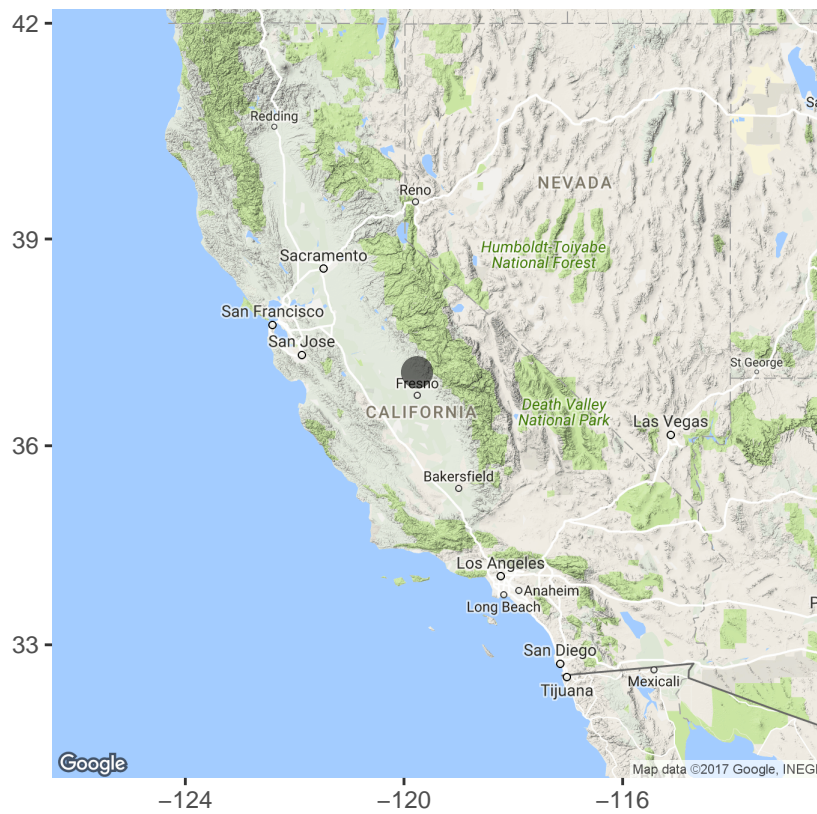
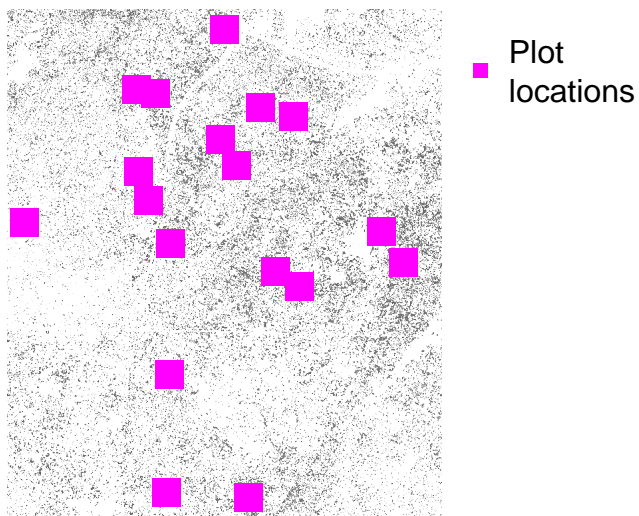Figure 3: ggmap of study area.

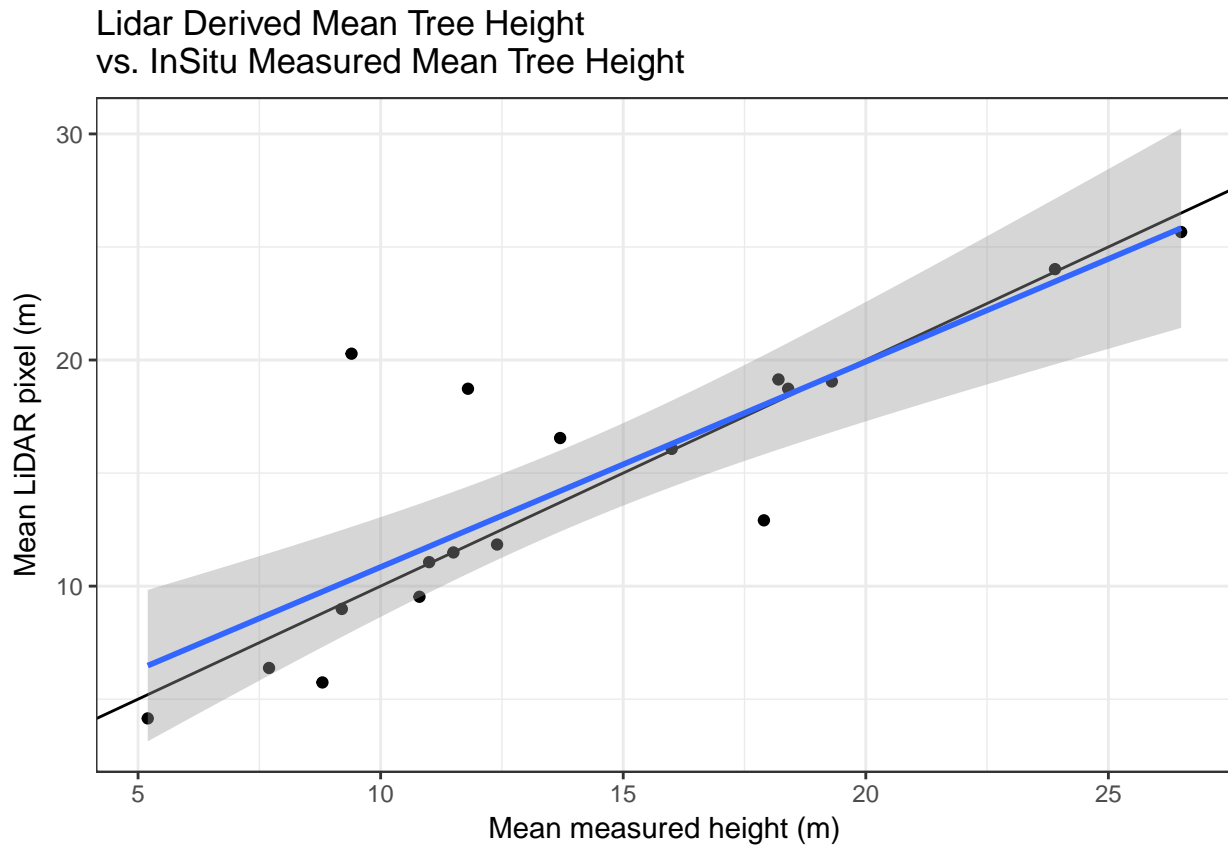## Study area plot locations

Figure 4: plots

Figure 5: final plot

```
<img src="~/Documents/Github/earthlab.github.io/images/course-materials/earth-analytics/week-5/buffer-c
<figcaption>The extract function in R allows you to specify a circular buffer
radius around an x,y point location. Values for all pixels in the specified
raster that fall within the circular buffer are extracted. In this case, we
will tell R to extract the maximum value of all pixels using the fun=max
command. Source: Colin Williams, NEON
</figcaption>
```

How different are the data?

View interactive scatterplot

View scatterplot plotly

View interactive difference barplot

View scatterplot differences

Additional Resources

The materials on this page were compiled using many internet resources including:
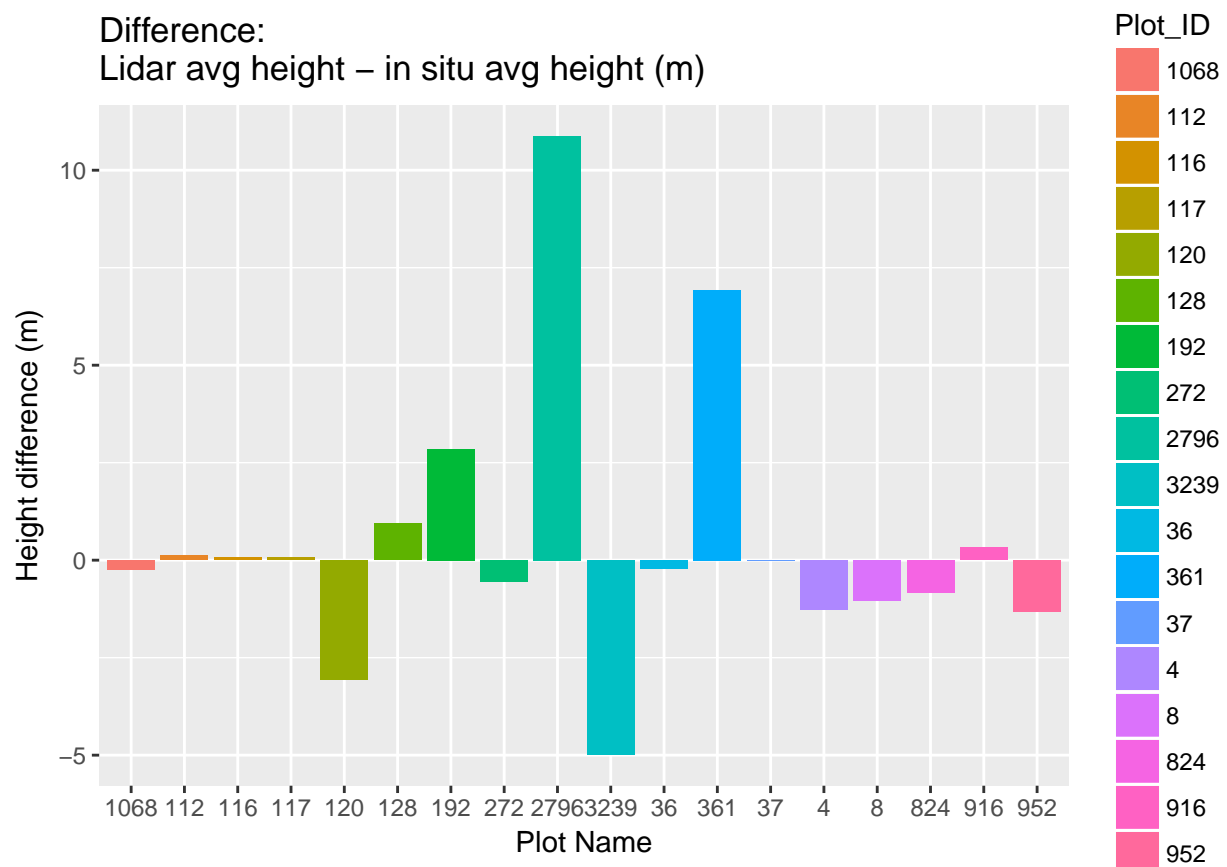
Figure 6: box plot showing differences between chm and measured heights.

- Pdf presentation on uncertainty in remote sensing
- An overview of key uncertainty terms
- A pdf that outlines uncertainty relative to land surface processes

QGIS imagery layer

- Code to add imagery to qgis via the python console: `qgis.utils.iface.addRasterLayer("http://server.arcgisonli` `<qgis._core.QgsRasterLayer object at 0x12739ee90>`