LiDAR, Machine Learning, and the Environment

Nate Dailey

November 2022

Abstract

LiDAR (Light Detection and Ranging) is a tool which uses laser pulses to spatially detect objects, similar to Radar's use of sound. LiDAR yields dense 3D point cloud data; the upside of this is its richness with information, and the downside is its difficulty to process. Machine learning is an extremely powerful tool to pair with LiDAR data and can produce highly accurate results. Models such as Support Vector Machines, Random Forests, and XGBoost have shown to be particularly useful, based on the literature surveyed in this paper. Machine Learning inspired approaches are also useful for compressing LiDAR data, which cuts down processing times. Accurate predictions from models such as these may be highly useful for LiDAR applications. Specifically, this survey paper will focus on applications of LiDAR in forestry.

Thesis: LiDAR is an extremely powerful tool for building computational interpretations of our world, and it is made even more so when combined with machine learning techniques.

1 Introduction

Over the past 20 years, LiDAR (Light Detection and Ranging) has completely changed how we collect quantitative information about forest environments. LiDAR's capability of providing rich data over extensive areas has made it much easier to study forest structure in a variety of domains, especially when compared to more traditional remote sensing techniques such as multispectral imaging and radar. LiDAR produces large datasets, which tend to be complementary to machine learning. Consequently, LiDAR and machine learning combine to form a highly useful tool in forestry.

How does LiDAR work? Similar to radar's use of sound pulses to measure distance, LiDAR maps the location of objects using light. Light pulses are radiated (e.g. from an aircraft), and after traveling through the air, are reflected (or absorbed) by objects and return back to the LiDAR sensor, which can then determine the position that the pulse reflected (that is, the location of the object).

Also, a single LiDAR pulse can produce multiple points (called returns), since some photons from the pulse may find their way through a meshed object

(e.g. a tree canopy), while others are immediately reflected, meaning the photons may find their way back to the sensor at different times. Some pulses and erroneous photons are lost, but a LiDAR apparatus can send tens of thousands of pulses per second, meaning that the area can typically be mapped with high precision.

LiDAR is particularly useful for mapping tree foliage, since the light pulses can travel through tree canopies and detect objects on forest floors. Having this property, LiDAR has been used extensively in archaeology to map areas obscured by dense forest canopies. For instance, in Caracol, Belize, LiDAR was used to study an ancient Mayan archaeological site which was hidden beneath a dense jungle canopy. [2]

What does LiDAR data actually look like? LiDAR data is represented by 3 dimensional point clouds (X, Y and Z floating point coordinates). Datasets tend to be extremely dense with points, so much that LiDAR data compression is an active area of research. [8]

LiDAR points can be filtered to create a variety of maps. For example, first return points can be filtered to create a Digital Surface Model (DSM), or last return points can be filtered (showing us the ground or rigid structures) in order to produce a Digital Elevation Model (DEM). This filtering process is typically executed in Geographical Information System (GIS) software.

A variety of metrics can be derived from LiDAR point cloud information, which can be used for a diverse set of purposes. For example, common metrics include tree height, canopy density, crown width, crown base height and crown volume. With further statistical analysis, one can extract characteristics such as fractional vegetation cover, aboveground biomass, timber volume, wood debris mass and canopy fuel volume (in a wildfire scope). [7]

LiDAR data is used for a variety of purposes. It can be used to identify tree species, such as in wildlife management and biodiversity research. It can also be used to study sustainable forestry and climate change impacts, providing insight into carbon stock [5], resource inventory, vegetation response to a changing climate, susceptibility to drought, insects and wildfire (for vulnerability estimation and use in fire modeling). LiDAR can even be used to study snowpack[1], which has important implications for drought conditions and other climate topics. [3]

Now that LiDAR has been explained in detail, this paper will focus on the application of machine learning on LiDAR data. Given machine learning's success with nonlinear problems, it has shown to be extremely useful in combination with LiDAR. For instance, models such as Support Vector Machines, Random Forests, and XGBoost have shown to be particularly useful.

2 Machine Learning vs. Traditional Models

Multiple Linear Regression (MLR) is the most common model for drawing information from LiDAR data. The advantage of this model is its clarity and simplicity. However, sometimes relationships in the data cannot be modeled linearly with high accuracy, leading us to look toward more complex tools. The

paper A Comparison of Machine Learning Regression Techniques for LiDAR-Derived Estimation of Forest Variables [4], compares a variety of machine learning tools to the more primitive MLR for classifying forest biomass attributes. Among the machine learning techniques are neural networks, support vector machines, nearest neighbor, Gaussian processes, Random Forest, and several types of regression trees.

The LiDAR was collected by an aircraft, flown on a path of 18 strips (back and forth, partially overlapping flight paths), over a forested area in the province of Lugo, in Galizia, Spain. From the height and intensity return values of the LiDAR data, features were extracted to be used as inputs for the machine learning models, alongside the point cloud data itself. Some examples of these extracted features are: percentage of first returns above 2m, number of returns above 2m, minimum return, maximum return, mean return, and so on. The output that the models attempt to predict includes the following forest attributes: stand crown biomass, stand stem biomass, and stand aboveground biomass. The correct labels used were recorded via manual field measurements.

Results are shown in Figure 1. For each algorithm, 100 iterations were used, and the result metric is R (the correlation coefficient). The best values achieved are in bold.

Algorithm	Maximum *R*	Mean *R*
MEAN-LM	0.717	0.711
MED-LM	0.716	0.711
LM	0.785	0.780
MDL-LM	0.786	0.781
RLM	0.796	0.791
kNN-LM	0.765	0.699
WkNN-LM	0.768	0.696
LWLR	0.687	0.562
MLR	0.916	0.913
IBk	0.959	0.845
SMO-p	0.976	0.944
SMO-g	**0.979**	**0.974**
M5P	0.891	0.891
GP-g	0.946	0.851
GP-p	0.970	0.929
RF	0.906	0.898
MLP	0.969	0.711

Figure 1: Averaged Maximum and Mean [4]

In the leftmost column in Figure 1, the upper 7 algorithms are variations of traditional linear models (LM), and the following 2 are linear regression models: Locally Weighted Linear Regression (LWLR) and Multiple Linear Regression (MLR). These are used as benchmarks for comparison with the machine learning models. Particularly, MLR, (mentioned previously) is a very useful reference point since it has been used widely with LiDAR historically.

The remainder of the algorithms are various machine learning models, and the specific implementations used were from the open source software WEKA.

IBk is an implementation of k-neighbor; Sequential Minimal Optimization (SMO) is a type of Support Vector Machine (with p and g suffixes representing polynomial and gaussian kernels, respectively); M5P is a M5 regression tree; GP is a Gaussian Process with standard (g) and polynomial (p) variations; RF is a Random Forest; and MLP is a Multilayer Perceptron.

We can see that SMO-g performed best, with a mean R value of 0.974 and a maximum R value of 0.979 (maximum represents the highest score over the multiple trial runs). SMO-g's mean R value of 0.974 is significantly higher than MLR's mean R value of 0.913, and so the authors conclude that the Support Vector Machine with a Gaussian kernel (SMO-g) is a better tool than the more traditional MLR for predicting forest attributes such as stand crown biomass, stand stem biomass, and stand aboveground biomass when used with LiDAR data. We can learn from this example that there is much potential for machine learning paired with LiDAR.

3 Classifying Crop Type with Machine Learning and LiDAR

In a 2020 study [6], researchers explored the usage of LiDAR with crop type classification, and compared its performance to imagery derived data (that is, data derived from images taken with cameras, a more traditional remote sensing technique).

Crop classification is highly important, since it helps governing bodies to monitor agriculture. Gathering crop data manually tends to be a labor intensive process, and accurate remote sensing classification would be extremely useful to this sector.

Machine learning models were used with LiDAR data, aerial (image) data (from a plane at 7500 ft) and satellite data (image) from Sentinel-2 satellite, and combinations of these sources. The results of these separate trials were compared in the study. The data was collected from an agricultural area in South Africa, with possible classifications of maize, cotton, groundnuts and orchards.

This study tested a variety of models. Most are from SKLearn's Python library, and more specifically, the following models from SKLearn were used: DT (Decision Tree), k-NN (K Neighbors), LR (Logistic Regression), NB (Naive Bayes), NN (Multi-layer Perceptron Neural Network), RF (Random Forest), SVM-L (Support Vector Machine - Linear), SVM-RBF (Support Vector Machine - Radial basis Function). XGBoost classification and Tensorflow's d-NN were also used. 100 iterations were performed, and the data was split by 70%/30% training/test sets.

We can see the results in Figure 2. A1, A2 and A represent trials of aerial imagery. S represents the Sentinel-2 satellite, L represents LiDAR, and hyphenated sets represent combinations of the data sources.

Classifier	Dataset									
	A1	**A2**	**S**	**L**	**A-S**	**A-L**	**L-S**	**A-S-L**	**Mean**	**Stdev**
d-NN	81	55.2	90.8	83.2	92.3	88.2	91.5	92.2	84.3	11.7
DT	72.2	46.1	81	82.3	86.2	84.7	90.2	90	79.1	13.6
k-NN	77.1	54.5	85.8	83.9	88.9	87.7	91.2	92.1	82.7	11.5
LR	73.2	44.5	85.3	84.9	91.6	86.8	92.2	92.9	81.4	15.2
NB	62.5	46.7	67.9	77.7	74.8	81.2	84.7	86	72.7	12.4
NN	81.2	56.5	88.2	86.3	92.7	89.8	92.8	93.4	85.1	11.5
RF	81.9	54.4	86.5	87.3	93.1	90.7	93.2	**94.6**	85.2	12.3
SVM-L	73.4	44	88.3	86.2	92.6	88.2	93.5	93.5	82.5	15.8
SVM RBF	72	50.5	75.9	83.4	87	86.6	89.8	90.1	79.4	12.5
XGBoost	81.3	56.1	86.3	**87.8**	91.9	91.3	93	94.1	85.2	11.7
Mean	75.6	50.9	83.6	84.3	89.1	87.5	91.2	**91.9**		
Stdev	5.8	4.8	6.6	2.8	5.3	2.8	2.5	2.4		

Figure 2: Accuracy of various machine learning models (leftmost column) used with different data sources.[6]

Overall, RF (85.2%), XGboost (85.2%), NN (85.1%), and d-NN (84.3%) performed the best (with percentages representing their mean performance across all data sources [10th column in Figure 2]).

The highest performing models differed for different data sources, but overall a combination of aerial, satellite and LiDAR (A-S-L) had the highest accuracy, with an average of 91.9% (bolded in Figure 1, and the RF model had the highest accuracy for A-S-L with an accuracy of 94.6% (bolded in Figure 2), with close contender XGBoost with an accuracy of 94.1% for A-S-L. Figure 3 below is an example of the imagery collected with several data sources.

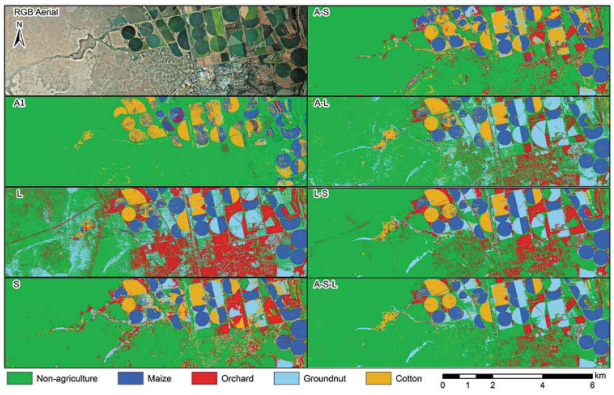

Figure 3: Visual representation of various data examined in study.[6]

In Figure 3, it makes sense that the combination datasets performed the

best, since the data sources can make up for each other's weaknesses. For instance, LiDAR accels at detecting height, but it is not as useful for detecting RGB/infrared differences between crops, and so the images collected from the aircraft and satellite made up for LiDAR's weaknesses in this case, perhaps. Since XGBoost had the highest accuracy for purely LiDAR data, at 87.8% (bolded in Figure 2), the study recommends XGBoost when just LiDAR is available.

All of this is to say that machine learning and LiDAR, especially when combined with other remote sensing techniques, can be extremely useful for practical problems such as crop type classification.

4 LiDAR Data Compression

A Common Challenge of Working with LiDAR data is its sheer size. With the most advanced sensors, millions of points can be collected per second, and even data representing small areas can reach into the gigabyte range. The paper *Real-Time LiDAR Point Cloud Compression Using Bi-Directional Prediction and Range-Adaptive Floating-Point Coding* [8] experiments with a new method for compressing LiDAR data using a machine learning-inspired approach.

More effective LiDAR data compression would be extremely useful for real-time applications, such as in autonomous vehicles, where the data must be processed and used instantaneously in order to make driving decisions. And more generally, more robust LiDAR compression is useful for all LiDAR applications, since cutting any computational costs is invaluable.

LiDAR compression involves quantizing (truncating mantissa values of floating point numbers) and removing redundant points. Standard MPEG compression algorithms tend to leave compression artifacts, which causes the data to lose precision. The model proposed in this study leverages range-adaptive floating-point coding (RAFC), an algorithm which uses machine learning principles to compress the data adaptively and minimize artifacts.

This study compared the new technique to two standard compression algorithms: MPEG G-PCC and MPEG Inter-EM. It was found that the new algorithm improved compression accuracy greatly at lower bit rates of input (bit rate represents the amount of data that each point holds, and so a higher bit rate corresponds to more dense LiDAR data). At higher bit rates, the 3 compression techniques produce more similar results, as can be seen by the Root Mean Squared Error in Figure 4.

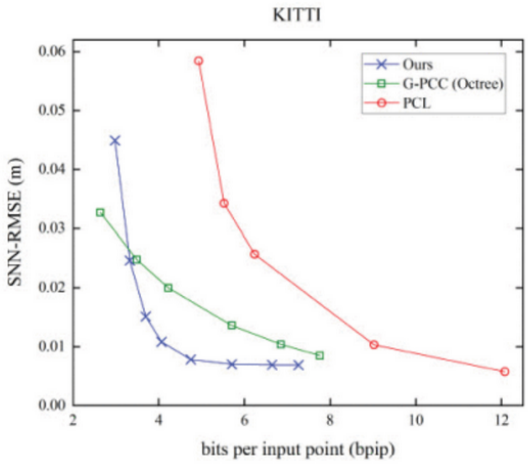

Figure 4: Accuracy of several compression algorithms [8]

This study shows that machine learning techniques can be used for LiDAR compression in order to boost efficiency of processing LiDAR data, an important step in expanding possible uses of LiDAR.

5 Discussion

In this survey paper, we have explored some possibilities of the synergy between machine learning and LiDAR. LiDAR data tends to be large and difficult to process, and the presence of machine learning allows us to achieve more accurate outcomes and speed up processing. As we saw in the case of using machine learning to predict forest biomass attributes [4] and classify crop types [6], we can use machine learning to gain valuable insights from LiDAR. We saw that models such as Support Vector Machines, Random Forests, and XGBoost have shown to be particularly useful. Also, we saw that machine learning-inspired adaptive approaches to LiDAR data compression can be extremely effective at minimizing data size while maintaining high fidelity. [8]

The optimizations in this paper point towards LiDAR being an extremely valuable tool, in the present and future. As researchers discover more ways to optimize the plethora of spatial data that it provides, we can see that LiDAR will become even more useful. Applied to fields such as forestry, LiDAR can help us to understand our changing planet without manual observation. It can help us more easily track wildfire risks, the health of our forests, snow cover, and watershed vitality–crucial metrics to understanding and supporting our environment. LiDAR is a powerful tool, and even more so when paired with modern computational techniques.

References

[1] Patrick D. Broxton, Willem J. D. van Leeuwen, and Joel A. Biederman. Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resources Research*, 55(5):3739–3757, 2019.

[2] Arlen F. Chase, Diane Z. Chase, John F. Weishampel, Jason B. Drake, Ramesh L. Shrestha, K. Clint Slatton, Jaime J. Awe, and William E. Carter. Airborne lidar, archaeology, and the ancient maya landscape at caracol, belize. *Journal of Archaeological Science*, 38(2):387–398, 2011.

[3] Michael J. Falkowski, Jeffrey S. Evans, Sebastian Martinuzzi, Paul E. Gessler, and Andrew T. Hudak. Characterizing forest succession with lidar data: An evaluation for the inland northwest, usa. *Remote Sensing of Environment*, 113(5):946–956, 2009.

[4] J. García-Gutiérrez, F. Martínez-Álvarez, A. Troncoso, and J.C. Riquelme. A comparison of machine learning regression techniques for lidar-derived estimation of forest variables. *Neurocomputing*, 167:24–31, 2015.

[5] Peter Hyde, Ralph Dubayah, Wayne Walker, J. Bryan Blair, Michelle Hofton, and Carolyn Hunsaker. Mapping forest structure for wildlife habitat analysis using multi-sensor (lidar, sar/insar, etm+, quickbird) synergy. *Remote Sensing of Environment*, 102(1):63–73, 2006.

[6] Adriaan Jacobus Prins and Adriaan Van Niekerk. Crop type mapping using lidar, sentinel-2 and aerial imagery with machine learning algorithms. *Geospatial Information Science*, 24(2):215–227, 2021.

[7] Kaiguang Zhao, Sorin Popescu, Xuelian Meng, Yong Pang, and Muge Agca. Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment*, 115(8):1978–1996, 2011.

[8] Lili Zhao, Kai-Kuang Ma, Xuhu Lin, Wenyi Wang, and Jianwen Chen. Real-time lidar point cloud compression using bi-directional prediction and range-adaptive floating-point coding. *IEEE Transactions on Broadcasting*, 68(3):620–635, 2022.