RANDOM FIELD MODELS FOR SPATIAL SMOOTHING OF AIRBORNE LIDAR TRANSECTS DATA

Amanda R. Coleman*[ab], Richard C. Olsen[a], Herbert K. H. Lee[b]
[a] Naval Postgraduate School, 1 University Circle, He-061A, Monterey, CA 93943-5006;
[b] University of California, Santa Cruz, 1156 High St, Santa Cruz, CA 95064

ABSTRACT

Light Detection and Ranging (LiDAR) is a form of remote sensing that utilizes laser scanners to produce a 3D point cloud of an environment by recording the number of laser pulse returns and measuring the backscattered energy as a function of time. LiDAR transect data were collected over the Monterey Peninsula and the Point Lobos Reserve. An experiment was conducted in the creation of a transect, a very high point density profile, by restricting the scan mirror with the initial goal of better understanding foliage penetration by LiDAR. Because of the high point density of the transect, the data were binned to create synthetic waveforms and to help reduce redundant points. However, the binning introduces sharpness in the data that distorts the typical wave shape in the synthetic transforms. A Bayesian Markov random field model captures the structure in the dataset and helps to offset the sharpness introduced during the binning. After fitting a Markov random field model using Markov chain Monte Carlo, classification methods were applied to distinguish objects in the landscape. These techniques should extend to true waveform data.

Keywords: Bayesian statistics, Markov random field, Support Vector Machine

1. INTRODUCTION

Light Detection and Ranging (LiDAR) is an active laser scanning, remote sensing method that can be used to map different characteristic structures across a region. In particular, LiDAR can be used to measure vegetation height, density, and urban areas attributes like roofs and roads, etc. across large swaths of land. Being able to capture

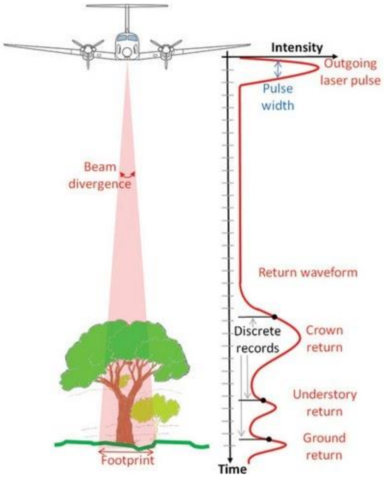

Figure 1-1 LiDAR data acquisition

information from a landscape without physically recording these characteristics by hand is a huge advantage of using LiDAR systems. LiDAR systems will record discrete returns or full waveforms.

As seen in Figure 1-1, one full waveform can be observed by the red line while the discrete returns are the points on the leading edge of the peaks. Discrete returns are obtained from the peaks in the recorded backscattered signal. The amount of signal, or energy, that returns is the intensity while each peak in a waveform often corresponds to an object like a tree, building, or ground. Because of its convenient compactness, LiDAR systems mostly only record discrete returns. Full waveforms require more storage space and processing techniques but can often provide more information than discrete returns alone. A sweeping, or circular motion of the scan mirror is the customary way to record LiDAR data. Operational LiDAR systems use a variety of scan patterns, including a transverse scan sweep. Uniquely, a collect was made in 2012 with the scan mirror turned off, in order to study a very high point density along a flight line. The transect was intended to help understand foliage penetration. The resulting data set was not compatible with traditional LiDAR processing techniques. Because of this, exploring statistical processing techniques, like a Markov random field for smoothing, became essential.

1.1 Previous Work

Since these transect data were collected in 2012, some initial analyses were conducted. Prof. Richard Olsen, presented the following at Optech's imaging and LiDAR Solutions Conference in 2013. The transect data are illustrated below, showing the returns plotted along the (roughly) 10 km flight line. The transect collect was compared to the traditional collect, and in particular a subset of the traditional data localized to within a few degrees of nadir. This 'synthetic' transect was created by searching the broad area collect within 20 cm of a transect point, not discriminating by scan angles, and extracting all returns associated with pulses intersecting with the transect. They used these transect data to discuss elevation pits over reflective paint, buildings under tree cover, trees over both sand and grass, skylights on a roof, and penetration through the tree canopy at this conference. Analysis showed that the degree of ground penetration between the transect and the synthetic transect were roughly the same (around 20%). Similarly, a large volume of duplicate points exists in the transect, roughly 96% had duplicate locations while only 51% had duplicate location and intensity.

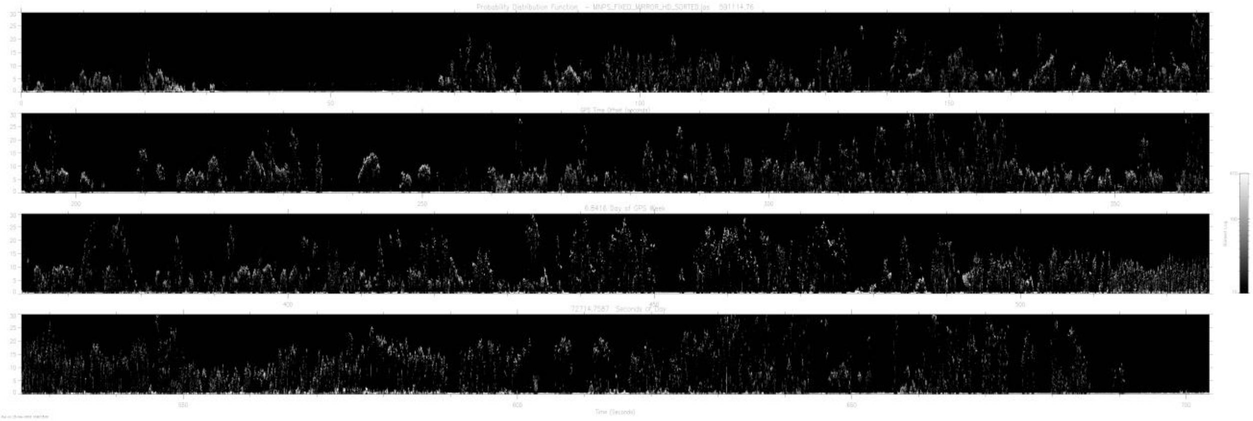

Figure 1-2 Data were collected with a 200 kHz PRF, at a rough altitude of 160 m.

Though the transect collection in concept dates back to the first days of laser scanner experimentation (e.g. Nelson et al, 1984), there are not many comparable works. Large spatial data were explored in papers by Andrew O. Finley and collaborators.[4] The previous work on LiDAR data focuses on discrete return data by applying spatial models directly to a point cloud, while the current project work looks at high density, binned, discrete data[1]. Finley and collaborators did an experimental collection that employed binning their discrete, point cloud data to create similar synthetic waveforms, or "pseudo-waveforms."[8] The data used here were a much higher point density due to the collection method and along a single track, which gives a different dimensionality than their gridded survey area. Where Finley binned 57 LiDAR samples together, this project explored using 50,000 and 10,000 samples, before settling on 1,000.[8]

1.2 Collection and Motivation

As part of the ongoing studies of foliage penetration, understanding how exactly light propagates through vegetation inspired a more unique approach.[2] A traditional LiDAR collect from a plane utilizes a sweep mirror that allows for collection over a large area per flight line in exchange for a lower point density and has the capacity to collect full waveform data of the terrain.[4] The transect data for this project contains points collected with the scan mirror turned off and thus does not contain full waveforms. With the scan mirror turned off for the collection of the transect data, the flight lines were collected at roughly 6000 - 8000 points/linear meter, along track, as opposed to 30 pulses per square meter in a traditional collection. Because of the vertical angle of the scanner for this collect as opposed to angled scanners from traditional collects, the conjecture was that this method could help inform how light permeates through the canopy.[2] A collection of this nature has never been attempted, therefore there are no established processing or analysis techniques for this type of data structure. Uniquely, the data were binned by the intensity of the high density transect data into essentially, synthetic waveforms. Because these data are along a single track and with such a high density, it follows that adjacent points are very spatially correlated and likely to be hitting the same object. This allows for different analysis methods than traditional collects. Because the intensity data can be displayed as an image, a Markov random field prior was chosen to model the data. The implementation of a Markov random field in this manner has not previously been explored. By using Bayesian modeling techniques, this paper will investigate the transect data to discern whether creating these synthetic waveforms will enhance analysis[6]. Lacking the software to implement complex LiDAR classification techniques common in the Remote Sensing discipline, a simple classification proof of concept will be performed to confirm the model's functionality.

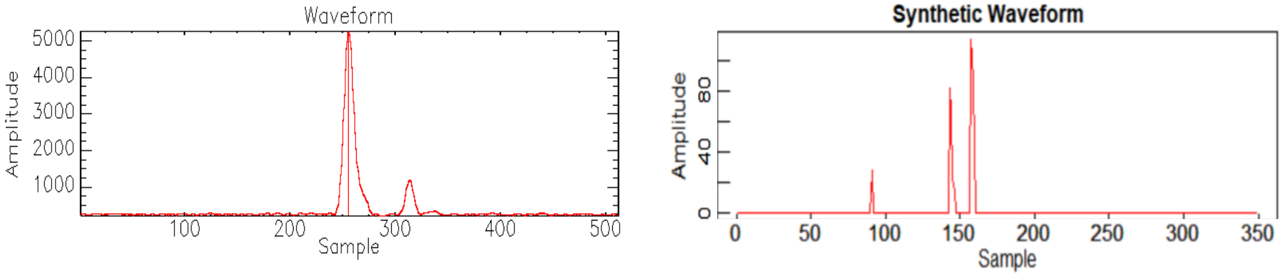

Figure 1-3 Example of Full Waveform from a different LiDAR collect

2. METADATA

The transect data come in the form of the LAS (.las) file format. These files contain point cloud information for the transect flight line. The raw data can be accessed by using a LAS reader like in IDL, MATLAB, or R. Before preprocessing, the transect LAS metadata files include X, Y, Z, Time, Intensity, Return Number, Number of Returns, Scan Direction Flag, Edge of Flight Line and Classification (empty), among a few others. However, most of the aforementioned parameters and the LiDAR transect proper prior to binning are not particularly interesting for several reasons, first and foremost being the sheer size of the data files. Another reason stems from the incredibly high point density that has a large percentage of redundant points. From here, the synthetic waveforms from the intensity values were created in order to reduce the redundancy. This will be the primary focus for analysis.

2.1 Preprocessing

Next, IDL was used to strip the .las files of the aforementioned variables. The data were binned by intensity values into histograms, using bin sizes of 1000 points. These histograms function as the synthetic waveforms of the data for intensity. Initially, exploratory data analysis began by using a bin size of 50,000. Various other sizes were also explored. A larger bin size did make the data size less cumbersome. However, the larger bin sizes created very sharp and rough looking waves that do not emulate true full waveform shapes very accurately. The choice to try this binning had several reasons. First and foremost, the choice to bin the data was due to the large amounts of duplicates from the high density collect. Secondly, the resulting synthetic waveforms resemble full waveforms and may potentially hold useful information for later classification. From the intensity and z vectors, this created a large matrix of 216,572 by 350 bands for the intensity synthetic waveforms. An individual synthetic waveform here is one sample of the 216,572 across all the bands as seen in Figure 2-3. Similarly, this binning process was used on the other variables, like Time, in order to line the data up properly.

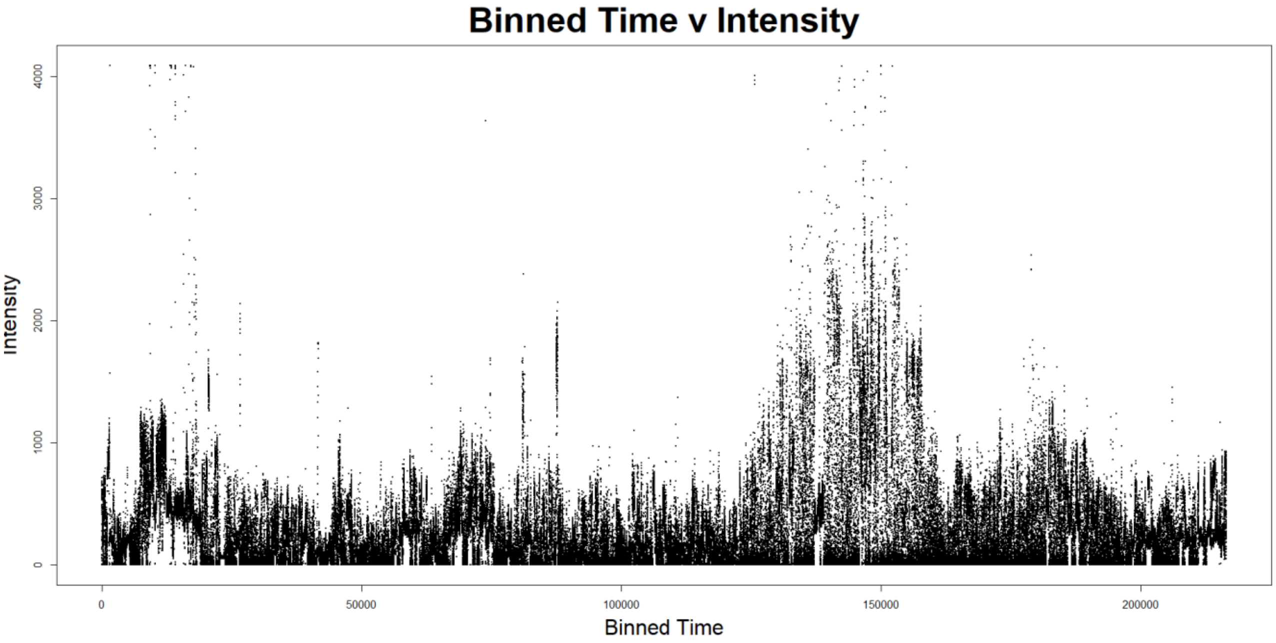

Figure 2-1 Entire flight line of data from band 1

2.2 Exploratory Data Analysis

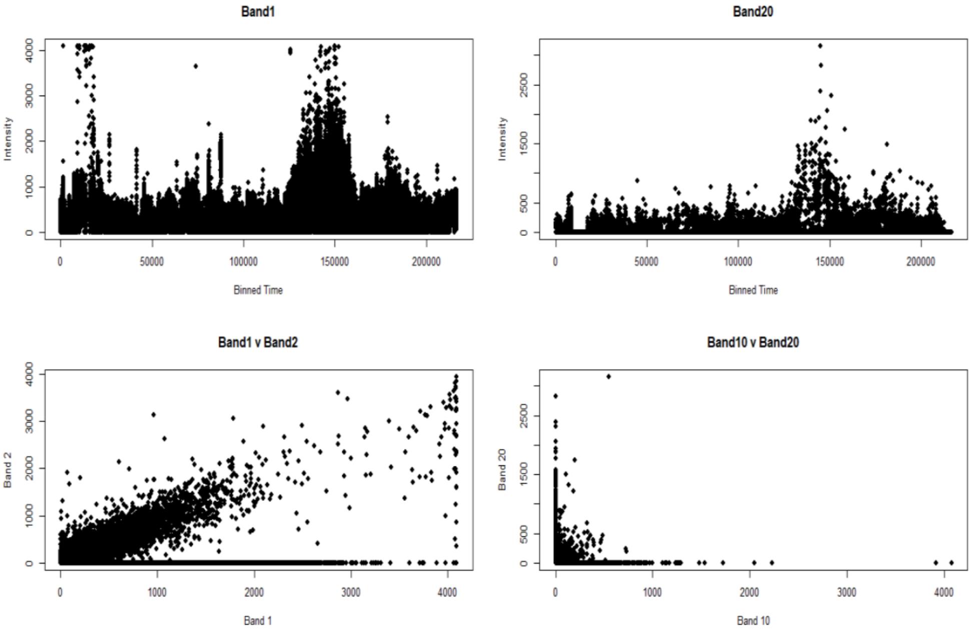

Figure 2-2 Correlation between bands and displaying how the number of samples with intensity decrease as the bands increase

While viewing the data, many extreme values appeared near the beginning and around the large peak near the latter half of the middle section, as seen in Figure 2-2. From the raw .las files, obvious noise and outlier points were manually removed using a point cloud viewer. Correlation between bands were also inspected. As seen in Figure 2-2, the bands are highly correlated. Similarly, the individual bands contain less information approaching the 350th band. The 350 bands come from how the LiDAR scanner records the returns. Nearly 100% of the data were contained in the first 350 nanoseconds following each initial pulse, 99.87% in the first 300, 97.18% in the first 200 nanoseconds, and 85.22% within the first 100 nanoseconds. For the later analysis, the data needed to be a much smaller subset for the Markov Chain Monte Carlo (MCMC) inference.[5] The subset used contains 140 bands and thus captures 92.04% of the data. Finally, individually waveforms were explored in order to confirm that the synthetic waveforms were behaving as expected. Seen in Figure 2-3, the waves do resemble full waveforms and are not behaving unusually.

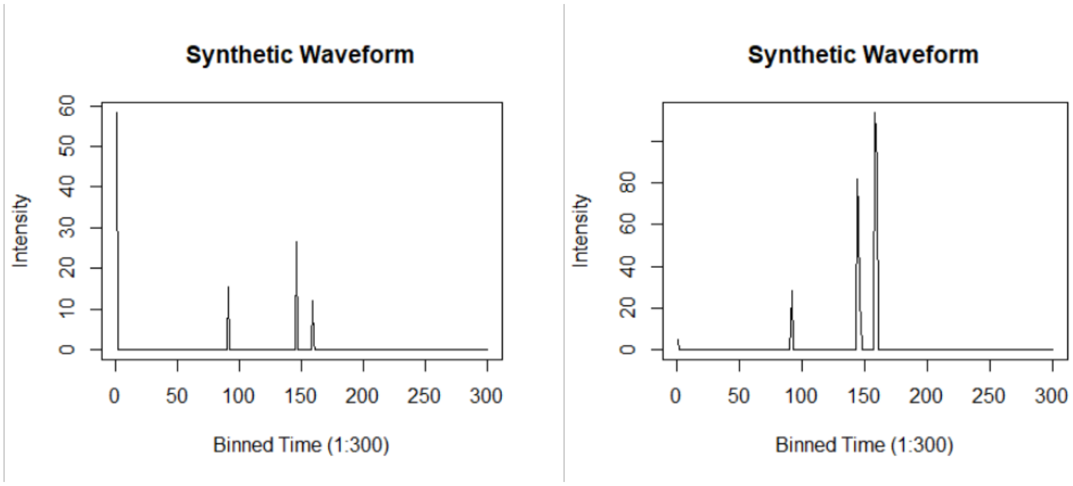

Figure 2-3 Different examples of the synthetic waveforms

Those pictured below are examples of the laser pulse hitting an object and returning. Though unlike true full waveforms, the synthetic waveforms pictured in Figure 2-3 are much less curved that the full waveform example in Figure 1-3.

2.3 Image Plots

With the data in a large matrix, an image plot will show different segments of the data. This image function in R rescales the data from zero to one. The x and y bounds are restricted using segments due to the massive size of the dataset. The image plots have time, or distance in the direction of the plane, for the x-axis and have the bands, or a z distance from the ground to the plane, on the y-axis. The image plots are then colored by the log intensity of the returns.

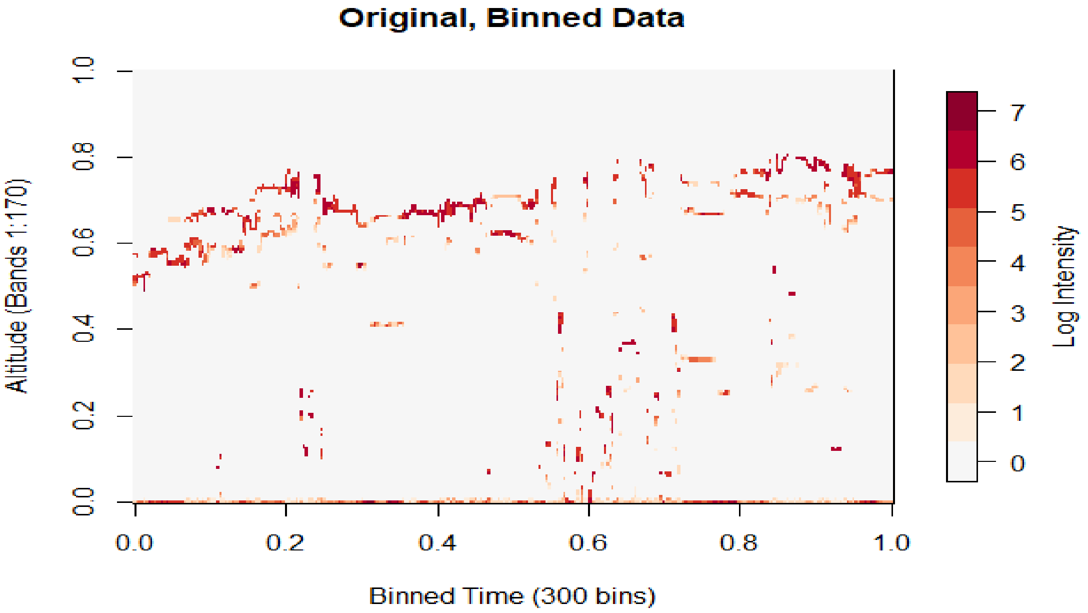

Figure 2-4 Image plot of several trees with overlapping canopies colored by intensity on the log scale.

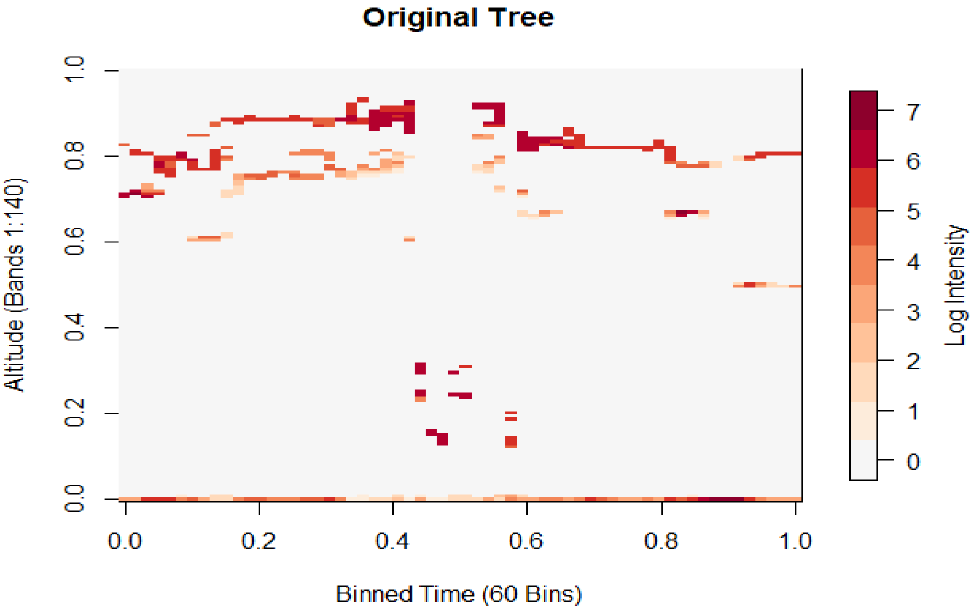

Figure 2-5 Image plot of an isolated tree, colored by log intensity, that will be used for the MCMC and analysis.

An image plot of multiple trees with overlapping canopy are depicted in Figure 2-4. From this image, understanding the difficulties in separating different structures in the data becomes obvious. The tree on the left (shown isolated in Figure 2-5) was used to fit the model as well as used for the analysis.

The tree in Figure 2-5 was selected for the analysis because this tree has clean sections of canopy, trunk, and ground. Specifically, there are clear returns, or intensity values, present for the tree canopy, trunk, and ground. If the backscattered laser signal does not encounter any object, it records as a zero, which is classified as air. From this image, the effect of the vertical laser and how, when there are several layers of returns, the layer closest to the laser for both the canopy and trunk have larger intensity values than those below is shown. The ground has both large and small intensity values, implying that the vertical laser can penetrate the canopy directly without refracting on the above canopy.

3. METHODS

The size of the data, despite the binning, still posed computational issues. Running an MCMC, even on a very small subset, would prove incredibly time and memory intensive. Despite these complications, convergence was achieved on the parameters when choosing appropriate sized subsets.

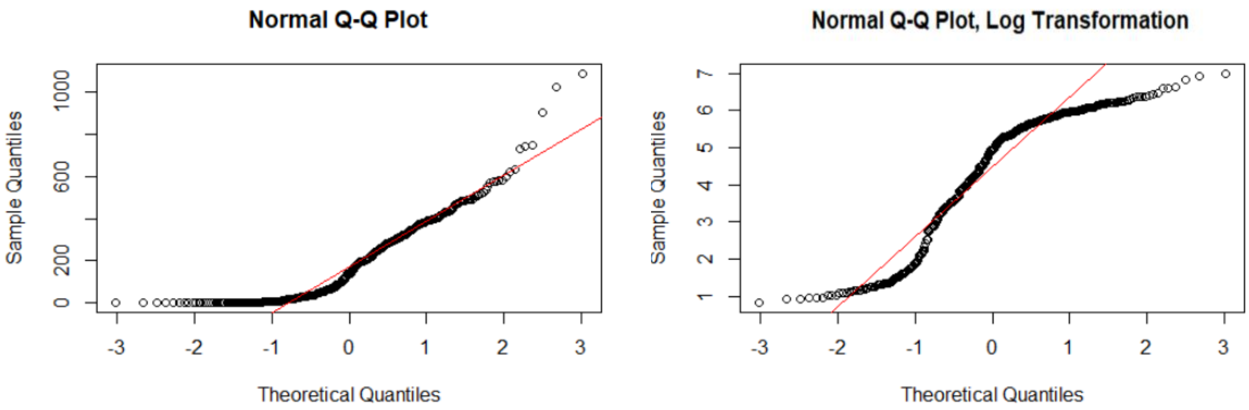

Figure 3-1 Normal Q-Q plots to check for normality

Before fitting a Markov random field model, the data needed to be transformed to make them closer to normal for the likelihood function. As seen in Figure 3-1, log transformation does remove the skewness. Despite making the tails a little too short, a log transformation makes the data much closer to normality than before the transformation.

3.1 Model Specification

The matrix data structure in tandem with a varying mean across different bands suggested using a Markov random field (MRF) as the prior for the intensity values. A random field can be used to model the structure between correlated points based on location or distance[7]. A Markov random field is a specific example of a random field that uses a multivariate Gaussian structure and represents the correlation structure based on locally defined neighborhoods. Conditioned on the neighborhood, a location's value is conditionally independent of points outside the neighborhood. Using a spatial model for smoothing fits the data structure of having high density LiDAR. A spatial model also implies that if one pixel is canopy, then adjacent pixels will also be canopy. This is especially true since across all the 350 bands, 98.7% of the entries are air: equal to zero. Once conditioned on the spatial dependencies using an MRF, the points can be treated as independent.

For implementing the model, the matrix of intensities needed to be transformed into a vector, y_i, that has length m. Since the intensities are highly skewed, a log transformation would bring the data much closer to normality. By adding 1 to each of the intensity values, y_i, and using a log transformation to compensate for the zero entries as seen in Figure 3-2, likelihood can be assumed as approximately normal and φ_i will be used to model the means for the log intensities.

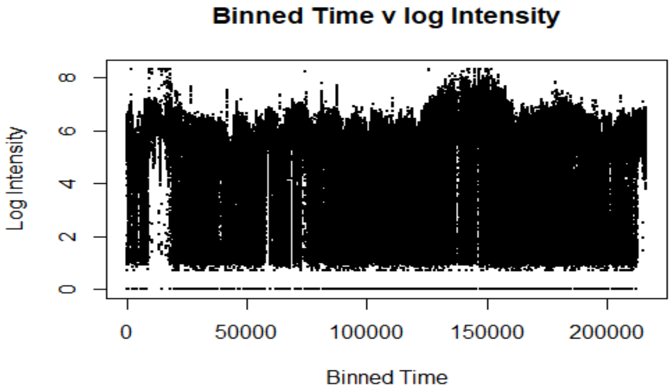

Figure 3-2 Plot of the data on the log scale. It can be seen in the plot above that the data do have a lot of zeros and therefore need to add a one to the y_i for $\log(y_i+1)$.

The likelihood is of the form:

$$L(\eta_i|\varphi_i,\sigma^2)\ \alpha\ \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\{-\frac{1}{2\sigma^2}(\eta_i-\varphi_i)^2\}$$

Where $\eta_i = \log(y_i + 1)$.

An inverse gamma conjugate prior on σ^2 was fitted where both a_{σ^2} and b_{σ^2} are fixed:

$$\pi(\sigma^2)\ \sim\ \Gamma^{-1}(a_{\sigma^2},b_{\sigma^2})$$

For a continuous-values process φ, defined over some m_1 x $m_2 = m$ lattice, a Gaussian Markov Random Field model can be used as a prior for θ, where theta is a precision parameter that controls the sparse. adjacency matrix, W.[3]

$$\pi(\varphi|\theta)\ \alpha\ \theta^{m/2}\ |W|^{1/2}\ \exp\{-\frac{1}{2\sigma^2}\varphi^TW\varphi\}$$

The adjacency matrix, W, is defined as $W = \{w_{ij}\}$, where n_i is the number of neighbors and w_{ij} defined as:

$$w_{ij}=\begin{cases}-1 & \text{if } i \text{ and } j \text{ are adjacent}\\ n_i & \text{if } i\ =\ j\\ 0 & \text{otherwise}\end{cases}$$

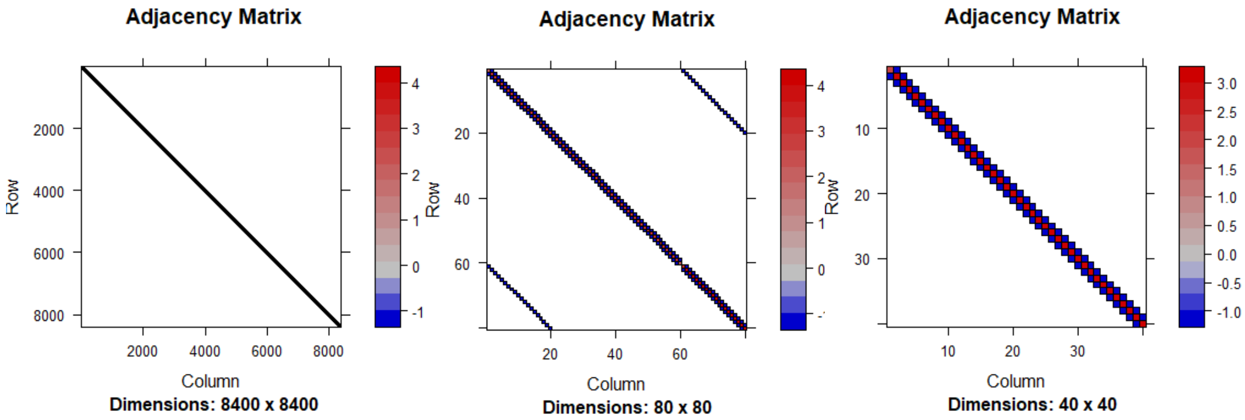

Figure 3-3 Adjacency matrix for the tree subset. Zoomed in from left to right, the neighborhood structure of the sparse matrix becomes more apparent.

Finally, there needs to be a prior on θ:

$$\pi(\theta) \sim \Gamma^{-1}(a_\theta, b_\theta)$$

Specifically, φ is the true values if the data were observed without noise or measurement error, θ reflects the correlation between neighboring pixels of φ, the 'smoothing' parameter, and σ^2 is the variance of the noise/measurement error in the data.

3.1.1 Hyperparameters

Selecting the hyperparameters for this model directly controls the precision and smoothing parameters. The hyperparameters were selected in order to vary the smoothing and the variance of noise in the data, θ and σ^2 respectively. A very small value of θ and a small value for σ^2 was desired. As seen in the collection of images below, the level of smoothing greatly varies within a small range for both the parameters. For the 60x140 subset for the MCMC, hyperparameters that provided a medium smoothing were selected from the below images. The hyperparameter values were fixed at $a_{\sigma 2} = 5$, $b_{\sigma 2} = 10$, $a_\theta = 8$, and $b_\theta = 1$.Using either the low or extreme smoothing combinations would either defeat the purpose of using an MRF to model the data or would eliminate too much information from the data.

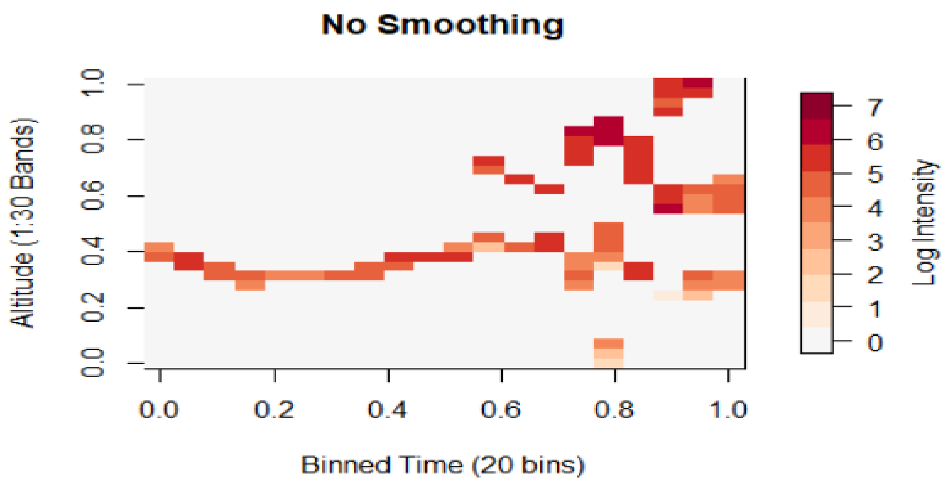

Figure 3-4. 20x30 subset using the first 30 bands. No smoothing

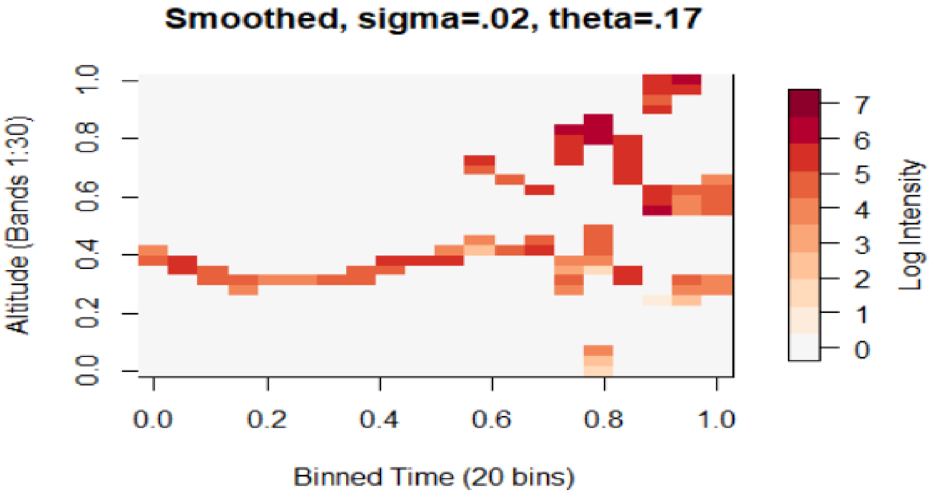

Figure 3-5. 20x30 subset using the first 30 bands. Mild smoothing

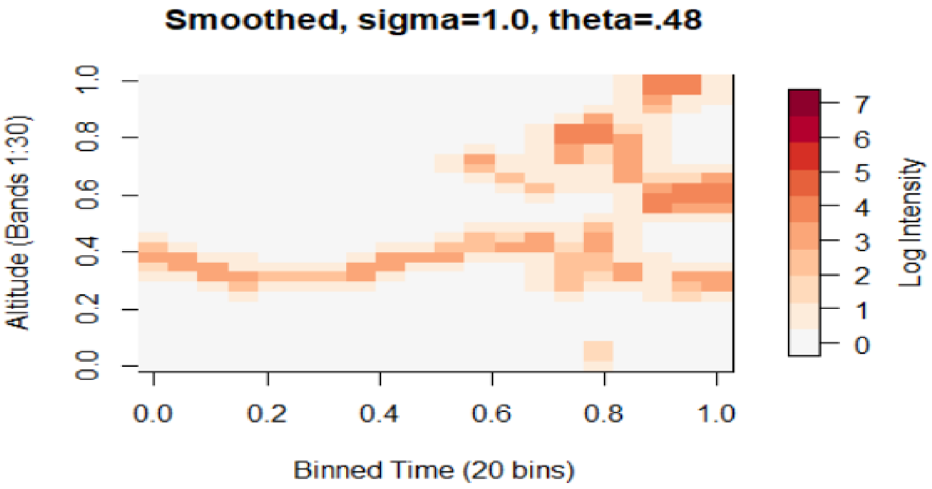

Figure 3-6. 20x30 subset using the first 30 bands. Medium smoothing

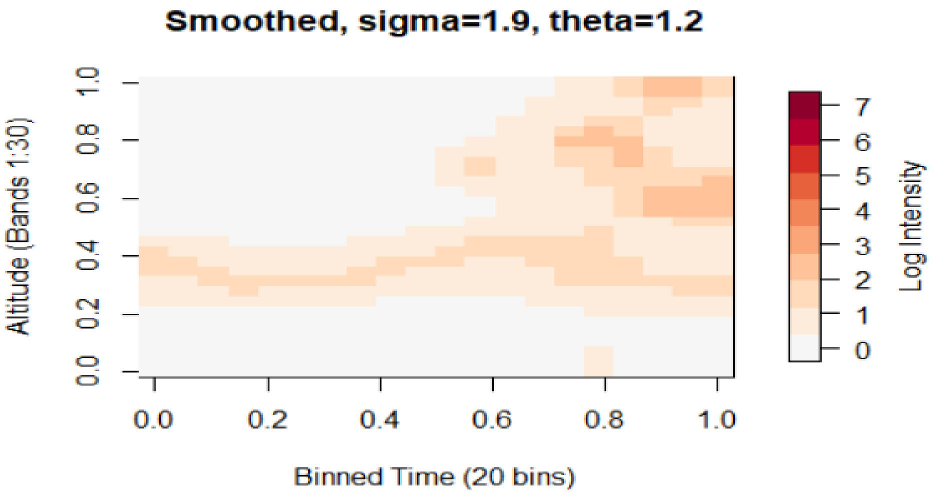

Figure 3-7. 20x30 subset. Extreme smoothing

3.2 Sampling and Full Conditionals

From the posterior distribution below, the lack of closed form implies that direct sampling cannot be used. Therefore, taking a Bayesian approach, full conditionals will be derived and used to sample from the posterior using a MCMC.

$$\pi(\theta, \sigma^2, \varphi|\eta_i) \; = \; L(\eta_i|\varphi_i, \sigma^2) \; * \; \pi(\sigma^2) \; * \; \pi(\varphi|\theta) \; * \; \pi(\theta)$$

$$\pi(\varphi|\theta, \sigma^2, \eta_i) \; \sim \; N(\; (\frac{\eta_i}{2} \; - \; \theta \; \sum_{i=1}^{m} w_{ij}\varphi_i)(\frac{1}{\sigma^2} + \theta w_{ii})^{-1} \; , (\frac{1}{\sigma^2} \; + \; \theta w_{ii})^{-1})$$

From the model structure, the full conditionals were derived for the unknown parameters in preparation for running an MCMC. The derived full conditionals are as follows:

$$\pi(\sigma^2|\varphi, \eta_i) \; \sim \; \Gamma^{-1}(a_{\sigma^2} \; + \; \frac{m}{2}, \; b_{\sigma^2} \; + \; \sum_{i=1}^{m}(\eta_i \; - \; \varphi_i)^2)$$

$$\pi(\theta|\varphi) \; \sim \; \Gamma(a_\theta \; + \; \frac{m}{2}, \; b_\theta \; + \; \frac{1}{2}\varphi^T W \varphi)$$

Where every w_{ii} provides the adjacency structure to each individual pixel of the image of intensities. The W matrix was constructed by defining $w_{ii} = n_i$ as the number of neighbors, setting w_{ij} = -1 if pixel i is adjacent to pixel j, and setting w_{ij} = 0 if i is not adjacent to j.[3]

Since all of the full conditionals are in closed form, the MCMC can be ran using only Gibbs steps. There was standard convergence for the parameters after 50,000 iterations. In terms of run time, the smaller 20x30 subset takes 10 minutes to run. However, the 60x140 subset, again capturing 92.04% of the data, takes over a week to run. This time constraint greatly impacts scaling up to larger subsets on a local machine without utilizing a GPU.

4. ANALYSIS

Next will be looking at model performance and classification.

4.1 Model Performance and Complications

From the 50,000 iterations, the mean was computed for each pixel from the log intensity parameter, φ. The means from the parameter, φ, can then be directly compared to the image plot of the tree before the smoothing. From the image plot in Figure 4-2, it is clear that the original and smoothed images are very related, showing that the model correctly fit the data.

Similarly, the colors are less intense in the smoothed image plot than the image plot of the original data. Furthermore, the data can also be compared analytically by comparing the summary output of each band for both the original and smoothed data sets in Figure 4-3. The quantiles in the summary are similar, but the smoothed data have quantiles slightly less, on average, than the original data.

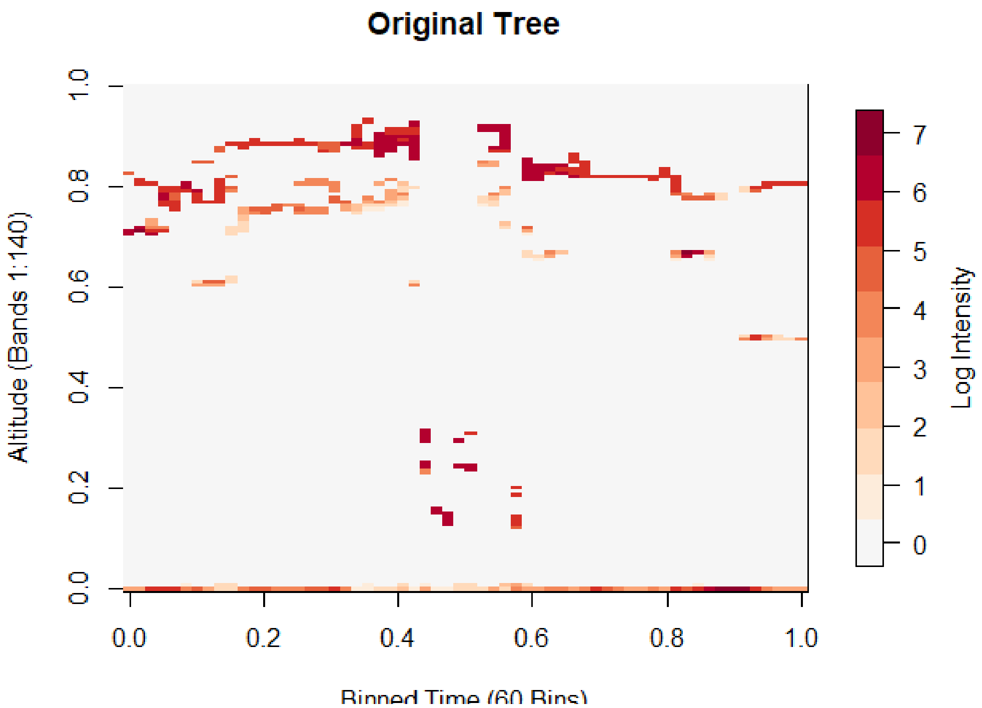

Figure 4-1 Tree before MCMC smoothing

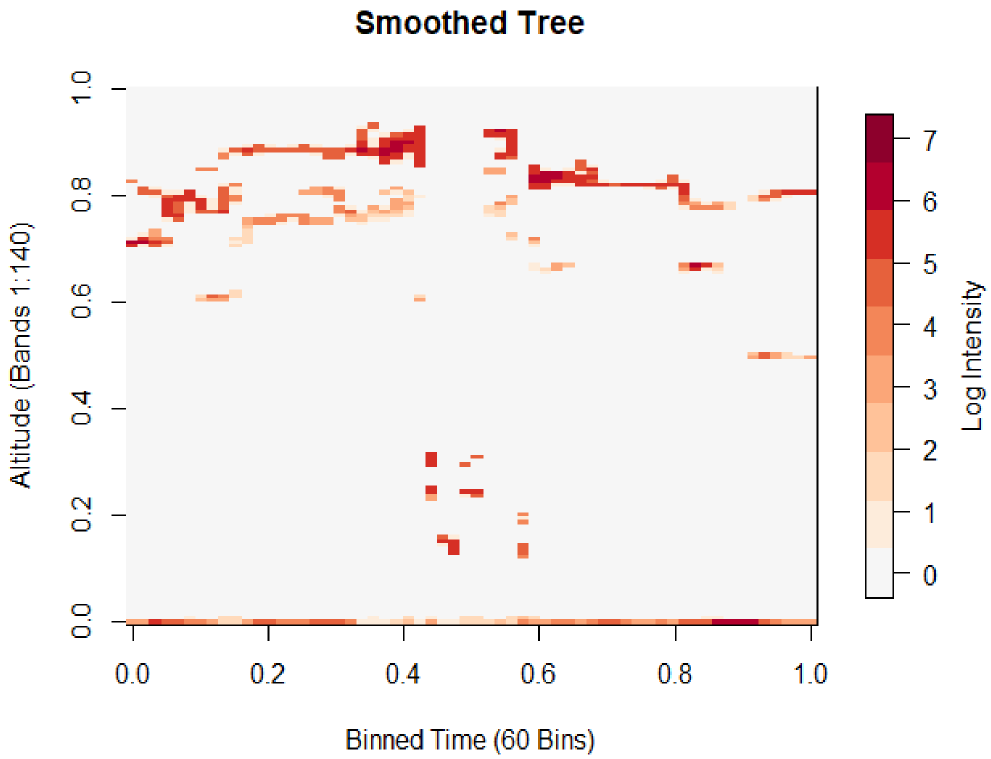

Figure 4-2 Tree after MCMC smoothing

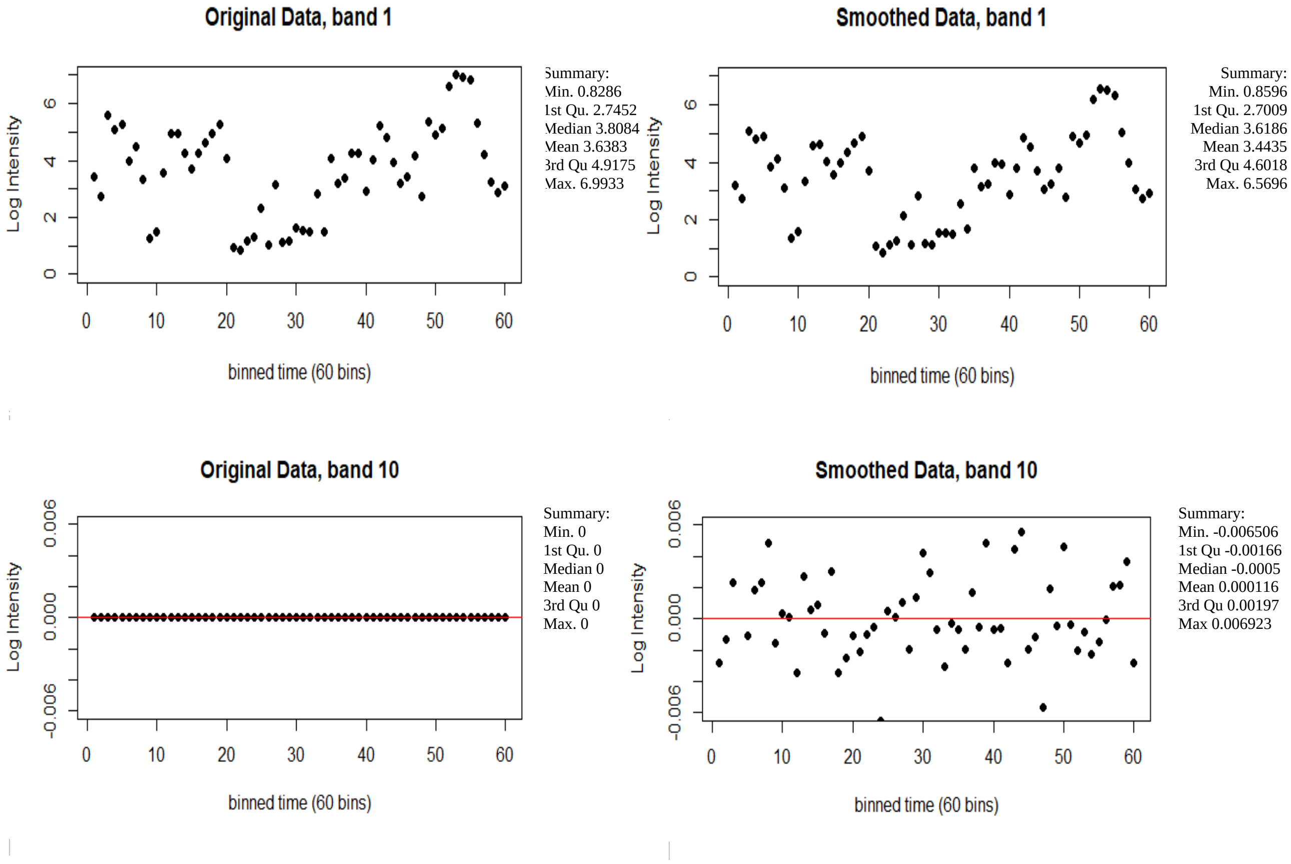

Figure 4-3 Before (left) and after (right) smoothing for band 1 and band 10 of the data.

Finally, the smoothing in the synthetic waveforms themselves can be assessed. Because of the binning process, the synthetic waveforms are very rigid and do not quite have the classic structure of a true full waveform. As seen in Figure 4-4, the synthetic, smoothed waveforms now more closely resemble the shape of a full waveform. This shows that a smoothing was achieved that would help eliminate the rigidity introduced by the binning process.

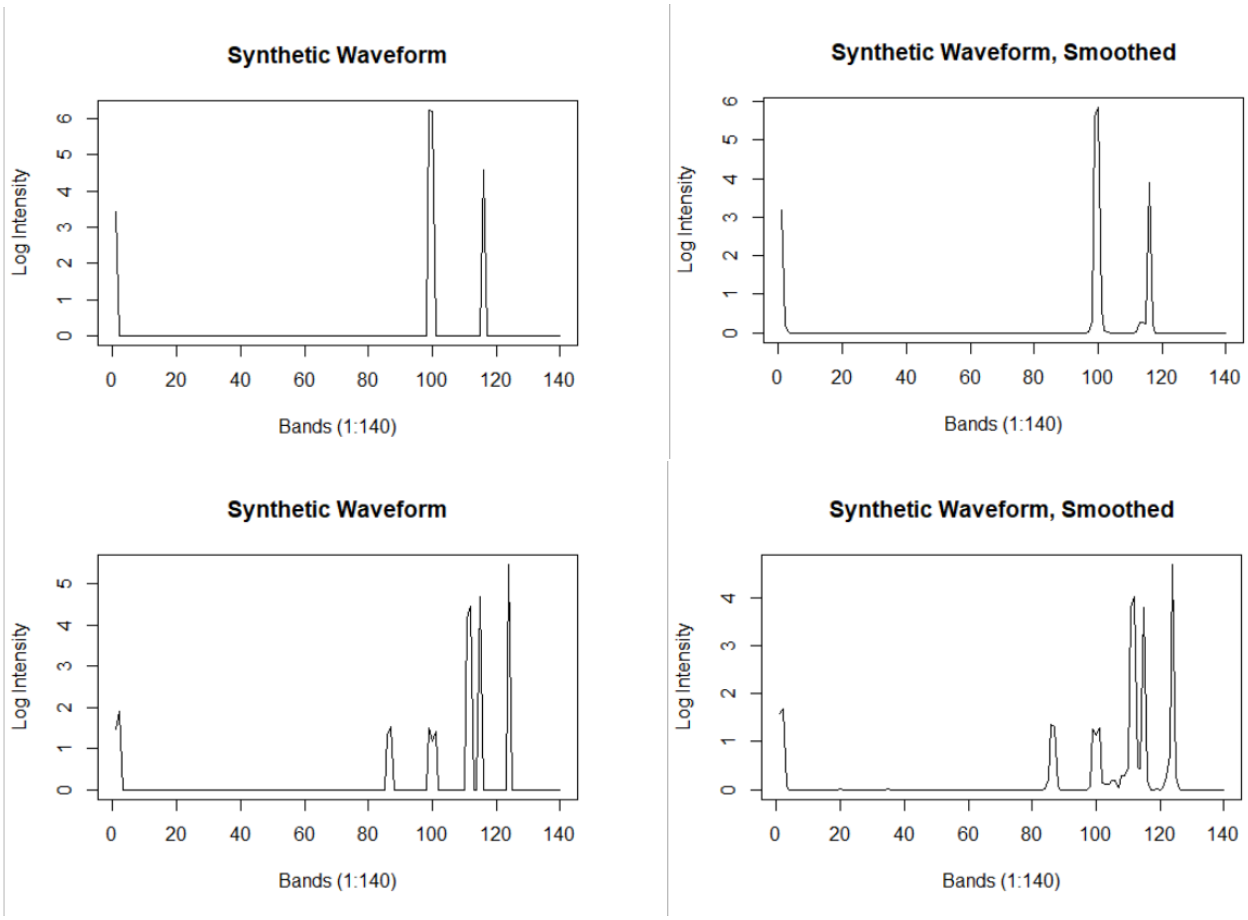

Figure 4-4 Before (left) and after (right) smoothing for the synthetic waveforms. The smoothed waveforms are clearly less rigid, especially around the base of each wave.

4.2 Classification

Given that physicists have their own software to classify data, below is a proof of concept on how the MRF can feed into classification methods. In order to classify the subset of the smoothed transect, another subset of equal size was created that also captured one tree. However, this tree was shorter than the tree used in the model. For better classification using these methods, training on multiple, isolated trees of various sizes would be recommended. Four classes were manually defined on the training set as ‘ground truth’. The classes were ‘air’, ‘ground’, ‘trunk’, and ‘canopy’. The smoothed intensities, time stamps, and bands were used to predict on these four categories. Because of the difference in tree size from the subsets, the classifiers had some unforeseen difficulty with differentiating the trunk and canopy.

Classification results from both a linear Support Vector Machine (SVM) and a multinomial logistic regression are consistent with the discrepancy as seen in Figures 4-5 and 4-6. Between these two classification methods, the multinomial logistic regression does, subtly, capture more of the canopy correctly than the linear SVM model. More specifically, compared to the manual ground truth, the linear SVM correctly classified 99.70% of the data and the multinomial logistic regression model correctly classified 99.64%. Using a Random Forest was also explored, but produced similar results to the linear SVM and the multinomial logistic regression.

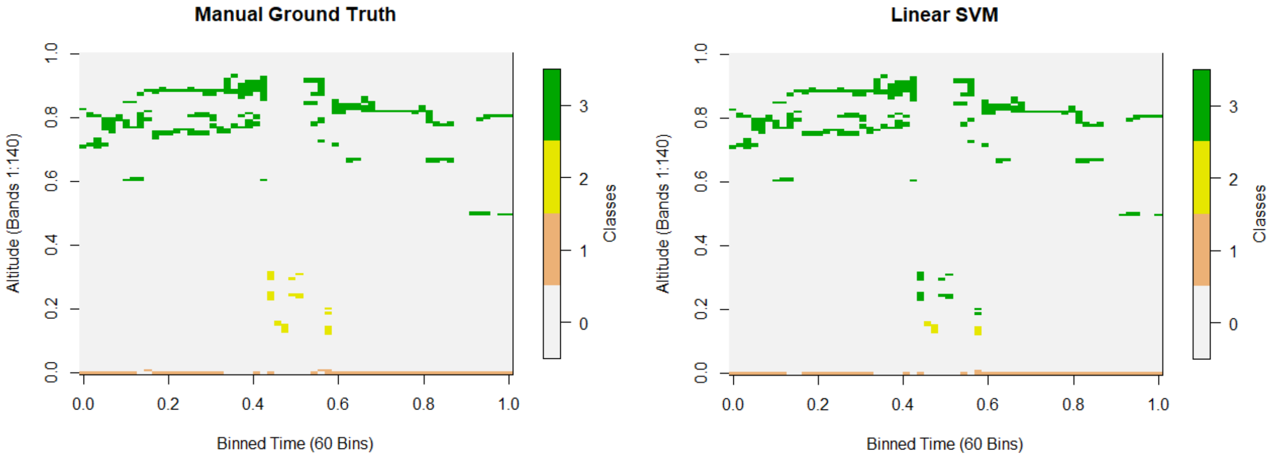

Figure 4-1 Both classified images are binned time (x-axis) by altitude (y-axis). Classification outcomes for using the Linear SVM (right) in MATLAB.

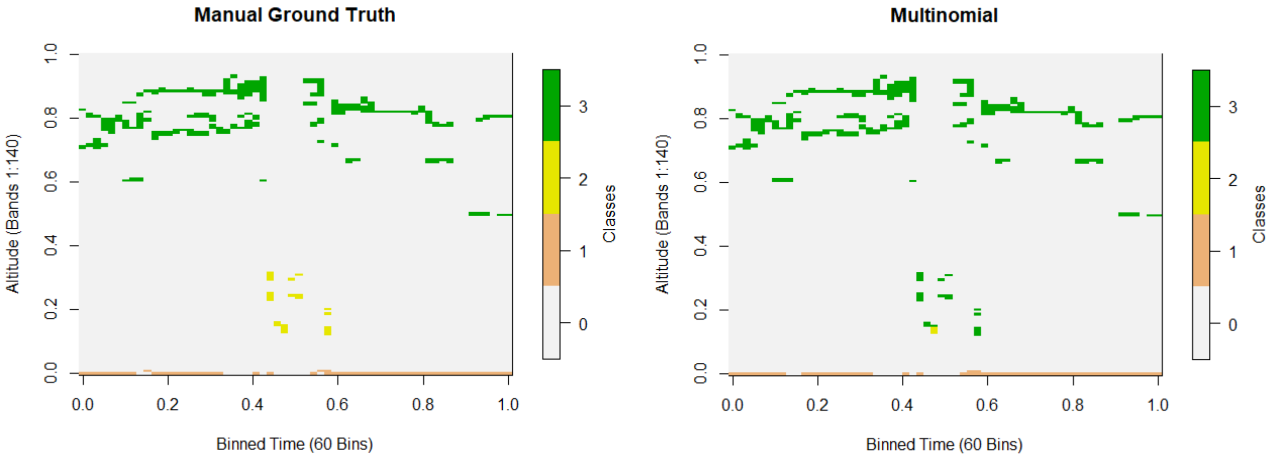

Figure 4-6 Both classified images are binned time (x-axis) by altitude (y-axis). Classification outcomes are from a Multinomial Linear Regression (right) in R.

5. DISCUSSION AND CONCLUSION

5.1 Conclusions

With the uniquely collected LiDAR data, an MRF model was successfully employed that could account for the spatial relationship between pixels. The applied MRF, through MCMC inference, was able to smooth the data that, in turn, reduced the rigidity in the synthetic waveforms introduced by the binning process. Despite reducing the size of the data by a factor of 1000 using binning, the large dataset proved difficult to process without accessing a GPU. Therefore, for the subset of a single tree, an MRF could be effectively used to smooth synthetic waveforms in order to more closely resemble full waveforms. Using the smoothed tree, a proof of concept was then applied to several classification methods.

5.2 Extensions and Future Work

Because of the uniqueness of the data and approach, many adaptations and extensions can be applied to this project for future works. Larger subsets could be processed using GPU and different models could be explored. Only with these larger subsets can more useful classification techniques be applied. Similarly, a more accurate ground truth data to classify subsets with more complexities than a single tree would also be needed. For more rigorous classification and consideration of more classes, MCMC results would need to be processed through established physics software. Finally, exploring whether this methodology could be extended to full waveform data remains to be seen. This process would entail restructuring true waveform data and using the MRF model to smooth these data with the goal of being able to better understand how light permeates through the canopy.

REFERENCES

[1] Finley, A., Banerjee, S. Zhou, Y., Cook, B., Babcock, C., “Joint hierarchical models for sparsely sampled high-dimensional LiDAR and forest variables”, Remote Sensing of Environment, 190, 149-161, (2017).

[2] Kim, A. M., Olsen, R. C. “Simulated lidar waveforms for the analysis of light propagation through a tree canopy.” In American Society for Photogrammetry and Remote Sensing, (2011).

[3] Lee, H. K. H., Higdon, D. M., Bi, Z., Ferreira, M. A. R., West, M., “Markov Random Field Models for High-Dimensional Parameters in Simulations of Fluid Flow in Porous Media,” Technometrics, 44(3), 230-241. (2002).

[4] Nelson, R., Krabill, W., MacLean, G. “Determining forest canopy characteristics using airborne laser data.” Remote Sensing of Environment. 15. 201-212. (1984).

[5] Olsen, R. C, Metcalf, J. P., “Visualization and analysis of lidar waveform data”, Proc. SPIE 10191, Laser Radar Technology and Applications XXII, 101910I, (2017).

[6] Robert, C. P., [The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation], New York: Springer Science Businesses, (2007).

[7] Rue, H. and Leonhard, H., [Gaussian Markov Random Fields: Theory and Applications], Boca Raton: Chapman and Hall/CRC, (2005).

[8] Taylor-Rodriguez, D., Finley, A. O., Datta, A., Babcock, C., Andersen, H. E., Cook, B. D., ... & Banerjee, S., “Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping,” Statistica Sinica, Volume 29(3), 1155-1180, (2019).

[9] Tso, B., and R. C. Olsen. “Combining Spectral and Spatial Information into Hidden Markov Models for Unsupervised Image Classification.” International Journal of Remote Sensing 26(10), 2113–33, (2005).