# LAND COVERS CLASSIFICATION BASED ON RANDOM FOREST METHOD USING FEATURES FROM FULL-WAVEFORM LIDAR DATA

Lian Ma, Mei Zhou, Chuanrong Li*

Key Laboratory of Quantitative Remote Sensing Information Technology
Academy of Opto-electronics, Chinese Academy of Sciences, 100094 Beijing, China - (malian, zhoumei, crli)@aoe.ac.cn

**KEY WORDS:** Random forest (RF), Full-Waveform LiDAR, Waveform decomposition, Skewness, Kurtosis, Land Covers Classification

**ABSTRACT:**

In this study, a Random Forest (RF) based land covers classification method is presented to predict the types of land covers in Miyun area. The returned full-waveforms which were acquired by a LiteMapper 5600 airborne LiDAR system were processed, including waveform filtering, waveform decomposition and features extraction. The commonly used features that were distance, intensity, Full Width at Half Maximum (FWHM), skewness and kurtosis were extracted. These waveform features were used as attributes of training data for generating the RF prediction model. The RF prediction model was applied to predict the types of land covers in Miyun area as trees, buildings, farmland and ground. The classification results of these four types of land covers were obtained according to the ground truth information acquired from CCD image data of the same region. The RF classification results were compared with that of SVM method and show better results. The RF classification accuracy reached 89.73% and the classification Kappa was 0.8631.

## 1. INTRODUCTION

Airborne Light Detection and Ranging (LiDAR) is a well-developed technique for 3D terrain modelling, which is finding increased usage in many different areas of application, such as environment monitoring, disaster assessment, and land covers classification. Compared to discrete return system, full-waveform LiDAR systems can record the entire backscattered waveform of the targets. The waveform features reflecting the properties of targets can be retrieved from the waveforms and are now extensively used for a large variety of land covers classification (Mallet, 2009). This paper aims to study the land cover classification using full-waveform LiDAR data.

As an important step in full-waveform LiDAR data processing, waveform decomposition techniques have progressed in recent years. Waveform features can be extracted through waveform decomposition. Land covers classification based on full-waveform features has been widely researched. The used features are usually intensity, width, pulse number, skewness and kurtosis (Mallet, 2008; Zaletnyik, 2010; Li, 2016).

In terms of classification, Waldhauser has developed models to automatically classify ground cover and soil types. Using the logic of machine learning, the advantages of supervised and unsupervised methods have been critically reviewed. The results showed that the supervised classifiers were preferred since they offer a higher flexibility (Waldhauser, 2014). The most commonly used are decision tree, Support vector machines (SVM) and Random Forests (RF) classification methods. Molijn has used decision tree to distinguish between four different types of terrain: snow, rock, ice and water based on width, reflectivity, saturation energy and kurtosis. However, the overall classification accuracy is only 74% (Molijn, 2011). Many researchers have investigated the waveform features for land covers classification using SVM

(Bretar, 2009; Tseng, 2015). Some scholars have compared RF classification and SVM classification, and showed that RF provides better classification accuracy (Dechesne, 2016). Chehata has studied different lidar features, multi-echo and full-waveform to classify urban scenes into four classes: buildings, vegetation, natural ground and artificial ground. The Random Forests classification using selected variables provide an overall accuracy of 94.35% (Chehata, 2009). Niemeyer has integrated a Random Forest classifier into a Conditional Random Field (CRF) framework for classifying urban LiDAR point clouds. (Niemeyer, 2014). Blomley has used Random Forest classifier for classifying airborne laser scanning data. Moreover, they have demonstrated that the consideration of multi-scale, multi-type neighbourhoods as the basis for feature extraction leads to improved classification results in comparison to single-scale neighbourhoods as well as in comparison to multi-scale neighbourhoods of the same type (Blomley, 2016).

In this paper, RF classification using full-waveform features, i.e. distance, intensity, FWHM (Full Width at Half Maximum), skewness and kurtosis was presented to predict the types of land covers as trees, buildings, farmland and ground. The classification result was compared with that of SVM method. The RF classification method shows higher classification accuracy. Moreover, the adaptability of the RF method was verified.

## 2. THEORY

### 2.1. Waveform processing method

Full-waveform LiDAR system records the entire backscattered waveform signals from targets, which are actually a sum of partial scattering response signals convolved with the scanner's system waveform. Thus it not only provides 3D point clouds, but also

obtains abundant information of the targets. By waveform processing, this information can be extracted. The waveform processing includes these parts: waveform filtering and waveform decomposition.

### 2.1.1 Waveform filtering

Before waveform decomposition, the noise of the waveforms needs to be removed. The widely used filtering methods include Wiener filter and Gaussian smoothing. However, the Wiener filter is very sensitive to the noise (Jutzi, 2006). Thus the raw waveforms are smoothed using a Gaussian filter (Brenner, 2003). For the Gaussian smoothing, it is crucial to select an appropriate width of the Gaussian kernel for each echo pulse reflected from the complex terrain. In this paper, the kernel width which is commonly described by the FWHM is defined via the standard deviation (sigma) of the transmitted pulse.

Figure. 1 shows the raw waveform and the filtered waveform of an echo. It can be seen that Gaussian filter has certain smooth effect on the raw echo waveform.
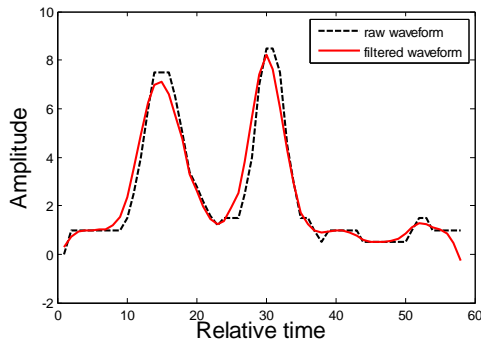


Figure 1.    Raw waveform and the filtered waveform.

### 2.1.2 Waveform decomposition

Since common laser transmitted pulse is modulated as Gaussian pulse, and the scattering of laser pulse for most targets can be approximated by a Gauss reflection, so the backscattered waveform component can be modelled as a Gaussian function. But when the targets are non-planar or an inclined plane, big errors would generate during waveform initialization and decomposition if Gaussian function is still used as the kernel function for modeling. Therefore, generalized Gaussian function was used for waveform modeling in this paper which can better represent the backscattered patterns from different targets (Zhou, 2015).

$$f(t) = \sum_{j=1}^{N} f_j(t) = \sum_{j=1}^{N} \hat{S}_j \exp(-\frac{|t - \mu_j|^{\alpha_j^2}}{2\sigma_j^{2}}) \tag{1}$$

Where, $f(t)$ is a waveform, $f_j(t)$ is the individual components of the waveform, $N$ is the components number, $\mu_j$ is the component position, $\sigma_j$ is the pulse width of individual components, $\alpha_j$ is the component shape factor.

The waveform decomposition procedures include initial parameters estimation and waveform fitting. In this paper, the first derivative is applied to estimate peak locations. The amplitude of the peak is extracted from the waveform at the peak location. Second derivative and first derivative are combined to calculate the

width of the components. In the fitting step, the classical Levenberg-Marquardt (LM) algorithm is used to solve the nonlinear curve fitting problem. Then the components parameters can be obtained (Duong, 2010).
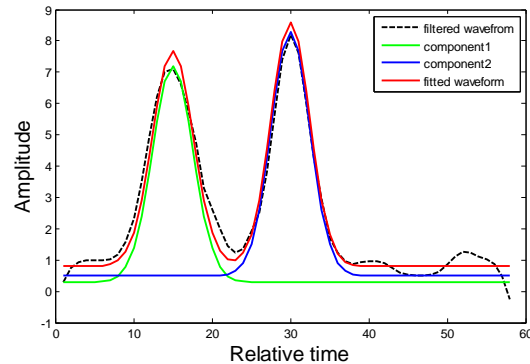


Figure 2.    Filtered waveform and decomposition results.

### 2.1.3 Waveform features extraction

Waveform features can be determined through the component parameters. The extracted features include distance, intensity, FWHM, skewness and kurtosis. Distance indicates the distance from laser transmitter to the target, which is determined by estimating the position of the waveform component. Ideally the peak position is considered as component position and the time lag is used to calculate the distance (Mallet, 2009). Intensity is a combination of emitted energy, distance, atmosphere attenuation and reflective capability of illuminated targets. In practice, the echo amplitude is most commonly regarded as intensity (Wagner, 2008). FWHM denotes the extension of waveform in the incident direction. It is closely related to geometry of targets and terrain slope (Duong, 2010). Skewness characterizes the degree of asymmetry of a distribution around its mean and kurtosis measures the relative peakedness and flatness of a distribution (Brenner, 2003).

Some factors, such as angle of incidence, atmospheric, range, surface characteristics, etc., have influence on the waveform features. To reduce such influence and further improve the effectiveness of waveform features for land cover classification, this work has made a comprehensive correction over some extracted waveform features. The detailed methodology was introduced in published article (Zhou, 2015).

## 2.2. Random Forest classification

Classification via Random Forest is performed in two phases. In the training phase the extracted features associated class labels is used to train a classifier.

In this paper, RF is chose as the classification method, since it is general and effective on many classification problems. Moreover, RF is robust to outliers and gives internal estimates of feature importance (Breiman, 2001).

RF operates by constructing a multitude of decision trees. Decision trees are built by choosing the most discriminative features in the feature vector, as a node to separate the training data according to their known class labels. As we know, Decision trees adapt to small variations and noise in the training data which results in overfit. Random forests overcome this issue by creating a large

number of decision trees. For each decision tree, random subsets of the training data are chosen and on each node only a random subset of the features are used. The parameters to be specified are the number of involved decision trees and the tree depth. In the experiments, we use a Random Forest with 20 trees and a tree depth of $\lfloor \sqrt{d} \rfloor$, where d equals the number of extracted features ,i.e. d=5. The tree depth is the number of features M, randomly chosen at each node, is considered as the single user-defined adjustable parameter. (Gislason, 2006).

For classification, each tree in the Random Forests gives a unit vote for the most popular class at each input instance. The label of input instance is determined by a majority vote of the trees. RF classifier aims at providing the highest prediction performance for the training data set (Oesau, 2015).

In the testing phase, the feature vector with unknown class labels is provided as input for classification by using the trained classifier.

Aside from classification, Random Forests provide measures of features importance based on the permutation importance measure which was shown to be a reliable indicator (Strobl, 2007). When the training set for a particular tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample set. These out-of-bag (OOB) data can be used to estimate the test accuracy and the permutation importance measure. This avoids the user to manually select relevant attributes.

## 3. EXPERIMENT AND ANALYSIS

### 3.1. Experiment data

The captured data of Miyun area, in Beijing, was used in this paper. The full-waveform data was acquired by the LiteMapper 5600 airborne LiDAR system. The average density of the point clouds was 4points/m$^2$. In this paper, a piece of experimental area containing about 385530 points was selected to study the RF classification. The CCD image of the selected area was shown in Figure 4 (a).

### 3.2. Experiment procedure

The flow chart of the experiment is shown in Figure 3. Firstly, the returned waveforms were smoothed using a Gaussian filter as mentioned above. Then waveforms were decomposed using LM algorithm. Features including distance, intensity, FWHM, skewness and kurtosis were extracted. The CCD image and pseudo color maps illustrating the values of extracted features, shown as Figure 4, were given and clearly illustrated the certain differences of each feature for different kinds of land cover types.

Secondly, 1000 features vectors for each typical land cover type were selected as the training data to train RF classifier according to the CCD image of the experiment region. Based on the 5 selected features, the Random Forests classification was run and variable importance was computed for each class. Underlying parameters have been fixed to M = 2 which means that two features are considered at each split and the number of trees was set experimentally to 20. Then the received waveforms reflected from typical land cover were divided into trees, buildings, farmland and ground using the RF classifier. Finally, the pseudo color classification image depicting the values of land cover types of Miyun area was generated and the results were evaluated.
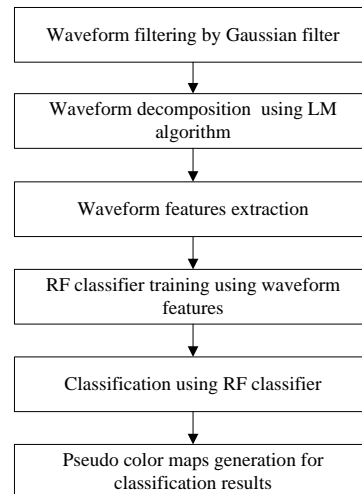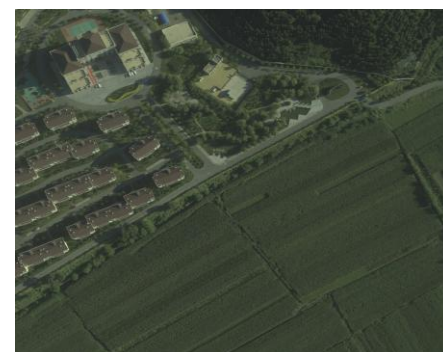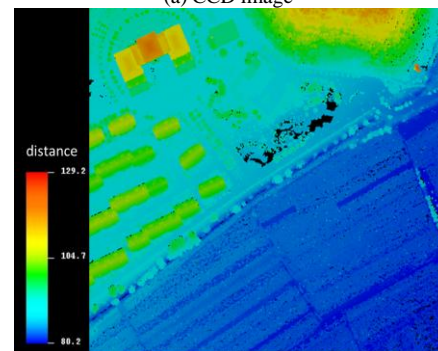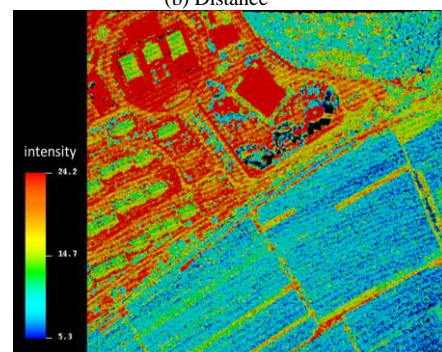


Figure 3. Flow chart for land cover classification of Miyun area based on full-waveform LiDAR data



(a) CCD image



(b) Distance



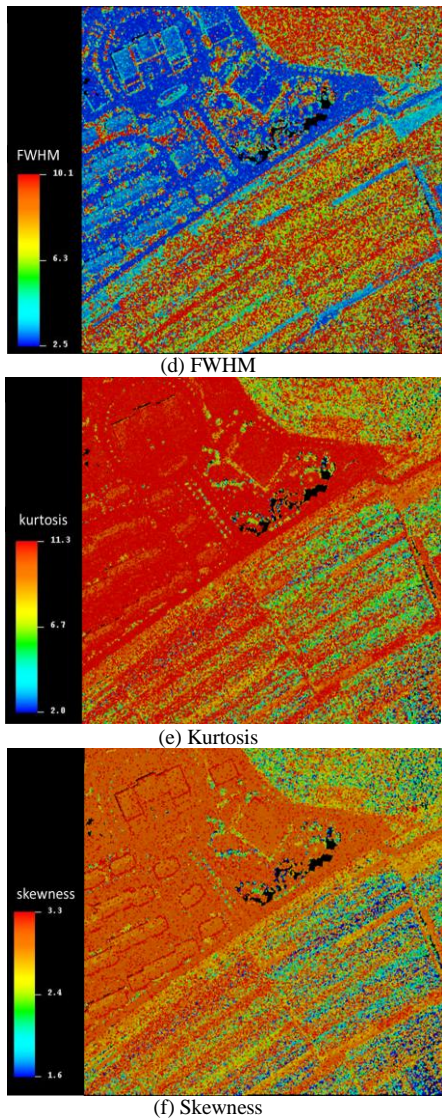(c) Intensity

(d) FWHM



(e) Kurtosis



(f) Skewness

Figure 4.    The CCD image and pseudo color maps of waveform features

### 3.3. Experiment results

The RF classification results are given in Figure 5. The green, red, wathet and yellow areas respectively represent trees, buildings, farmland and ground. The classification results of RF were compared with that of SVM, which has been studied in our previous work (Zhou, 2016). The SVM classification results are shown in Figure 6. It can be seen that by using the features we can effectively distinguish different types of land covers based on RF method. While for SVM, many building points are incorrectly classified as trees, as the purple rectangle shows; many farmland points are incorrectly classified as trees, as the orange rectangle shows. From the figures, we can see that the RF has better classification results than SVM.
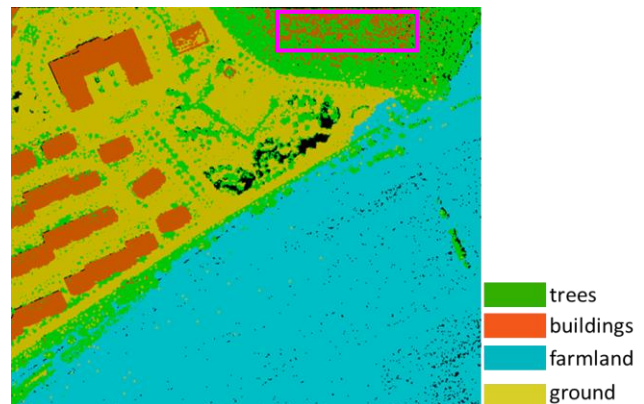


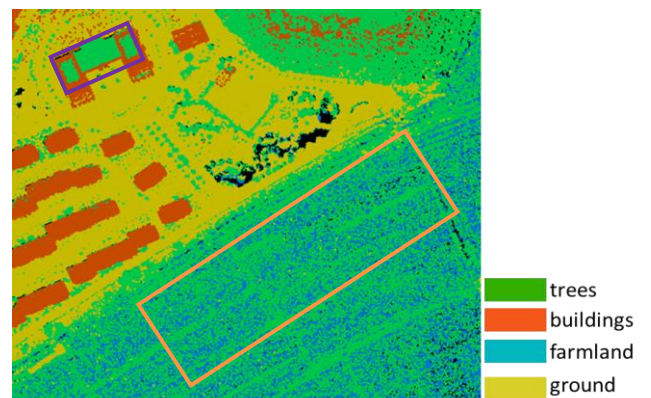Figure 5.    RF classification results of Miyun area



Figure 6.    SVM classification results of Miyun area

6400 instances of these four land cover types, except the training data, were selected to calculate the classification accuracy. The ground truth information was acquired manually according to CCD image, as shown in Figure 4 (a). The confusion matrix for the RF and SVM classification results of these four land cover types of Miyun area were obtained and shown in Table 1 and Table 2, respectively. The overall classification accuracy for RF and SVM method are 89.73% and 0.8631, respectively. While the classification Kappa for RF and SVM method were 82.17% and 0.7623, respectively. It can be seen that the RF has higher classification accuracy.

| Truth / Classification | Trees | Buildings | Farmland | Ground | Total |
|---|---|---|---|---|---|
| Trees | 1598 | 143 | 0 | 264 | 2005 |
| Buildings | 2 | 1209 | 0 | 0 | 1211 |
| Farmland | 0 | 0 | 1600 | 0 | 1600 |
| Ground | 0 | 248 | 0 | 1336 | 1584 |
| Total | 1600 | 1600 | 1600 | 1600 | 6400 |

Table 1. Confusion matrix of the RF classification results

| Truth / Classification | Trees | Buildings | Farmland | Ground | Total |
|---|---|---|---|---|---|
| Trees | 1567 | 695 | 123 | 23 | 2408 |
| Buildings | 28 | 638 | 0 | 0 | 666 |
| Farmland | 0 | 0 | 1477 | 0 | 1477 |
| Ground | 5 | 267 | 0 | 1577 | 1849 |
| Total | 1600 | 1600 | 1600 | 1600 | 6400 |

Table 2. Confusion matrix of the SVM classification results

In order to verify the adaptability of the RF method, another piece of data in Miyun area was classified using the same training model. It shows that the classification results are generally good, but there are also some misclassification between trees and buildings, trees and farmland. Therefore, the RF method has certain adaptability but should be further improved.
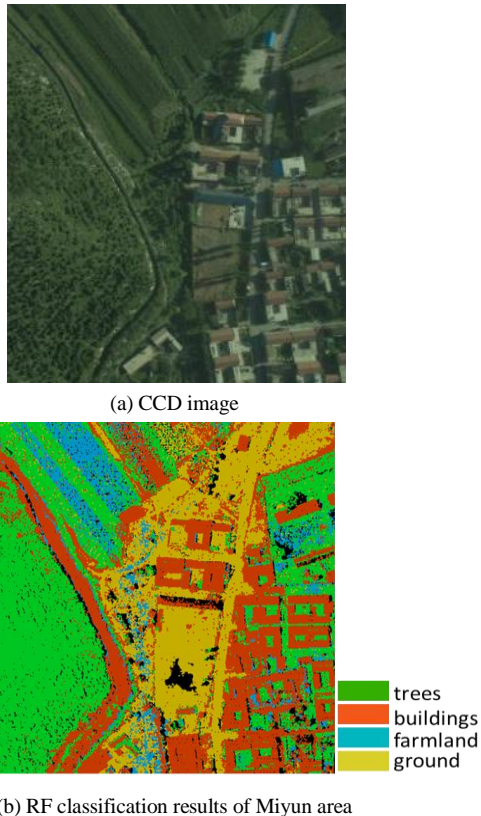


(a) CCD image



trees
buildings
farmland
ground

(b) RF classification results of Miyun area

Figure 7.     CCD image and RF classification results of another area

### 3.4. Analysis

From the above results, we can see that the RF method has better classification results than that of SVM method. However, there are also some classification errors for RF method. From Figure 4 we can see that some area on the upside of the Figure was "trees" in fact; however it was classified to be "buildings", as the pink rectangle shows in Figure 5. As shown in the second row and the second column in Table 1, we can also see that the most confusion in prediction by RF classifier was between "buildings" and "trees". This was possibly resulted from the similar distance of "buildings" and "trees". Prediction errors were also generated from "trees" and "ground", as shown in the second row and the fifth column in Table 1, which was because that the ground points are beside the trees points, and the area of ground points is small, when selecting the test data, some tress points are selected to be ground points.

### 4.    CONCLUSIONS

In this paper, the returned waveforms were smoothed using a Gaussian filter and waveform decomposition was implemented. Then waveform features were extracted. RF classifier was generated to classify the types of land covers in Miyun area as trees, buildings, farmland and ground. The RF classification results were compared with that of SVM method and show better results. The RF classification accuracy reached 89.73% and the classification Kappa was 0.8631. In the future work, for obtaining better classification results, more geometric 3D and 2D features will be extracted and correlation between the features will be analysed. The adaptability of classification method will be further improved.

### REFERENCES

B. Jutzi, and U. Stilla, 2006. Range determination with waveform recording laser systems using a Wiener Filter. *Isprs Journal of Photogrammetry & Remote Sensing*, 61, pp.95-107.

Blomley, R., Jutzi, B., & Weinmann, M., 2016. Classification of airborne laser scanning data using geometric multi-scale features and different neighbourhood types. *Isprs Annals of Photogrammetry Remote Sensing & Spatial Informa, III-3*, pp.169-176. Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp. 5–32.

Brenner, A. C., H. J. Zwally, C. R. Bentley, B. M. Csatho, D. J. Harding, M. A.Hofton, J. B. Minster, L. A. Roberts, J. L. Saba, R. H. Thomas, and D. Yi, 2003. Geoscience laser altimeter system algorithm theoretical basis document: Derivation of range and range distributions from laser pulse waveform analysis. *Algorithm Theoretical Basis Documents (ATBD)*, 92pp.

Bretar, F., Chauve, A., Bailly, J. S., Mallet, C., Jacome, A., 2009. Terrain surfaces and 3-d land cover classification from small footprint full-waveform lidar data: application to badlands. *Hydrology & Earth System Sciences Discussions*, 13(8), pp. 1531-1544.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne LiDAR feature selection for urban classification using random forests. *International Archives of thePhotogrammetry, Remote Sensing and Spatial Information Sciences,* 39 (Part 3/W8), pp.207–212.

Dechesne, C., Mallet, C., Bris, A. L., Gouet, V., and Hervieu, A., 2016. Forest stand segmentation using airborne lidar data and very high resolution multispectral imagery. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* pp.207-214.

Duong, H. Processing and application of icesat large footprint full waveform laser range data. Doctor thesis of the Delft University of Technology, Netherlands, 2010.

Gislason, P., Benediktsson, J. and Sveinsson, J., 2006. Random forests for land cover classification. *Pattern Recognition Letters*, 27 (4), pp. 294–300.

Li, Q., Ural, S., and Shan, J., 2016. Decomposing LiDAR waveforms with nonparametric classification methods. *IGARSS*

*2016 - 2016 IEEE International Geoscience and Remote Sensing Symposium*, IEEE. pp.5573-5576.

Mallet, C., Bretar, F., 2009. Full-waveform topographic lidar: state-of-the-art. *Isprs Journal of Photogrammetry & Remote Sensing*, 64(6), pp.1–16.

Mallet, C., Soergel, U. and Bretar, 2008. F. Analysis of full-waveform lidar data for classification of urban areas. *Photogrammetrie Fernerkundung Geoinformation*, 74(5), pp. 337-349.

Molijn, R. A., Lindenbergh, R. C., & Gunter, B. C., 2011. Icesat laser full waveform analysis for the classification of land cover types over the cryosphere. *International Journal of Remote Sensing, 32*(23), pp.8799-8822.

Niemeyer, J., Rottensteiner, F., & Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. *Isprs Journal of Photogrammetry & Remote Sensing, 87*(1), pp.152-165.

Oesau, S., Lafarge, F., and Alliez, P., 2015. Object classification via planar abstraction. *Isprs Annals of Photogrammetry Remote Sensing & Spatial Informa*, *III-3*, pp. 225-231.

Tseng, Y.H., Wang, C.K., Chu, H.J., Hung, Y.C., 2015. Waveform-based point cloud classification in land-cover identification. *International Journal of Applied Earth Observation and Geoinformation*, 34(1), pp.78-88.

W. Wagner, J. Hyyppä, A. Ullrich, H. Lehner, C. Briese, S. Kaasalainen, 2008b. Radiometric calibration of full-waveform small-footprint airborne laser scanners. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, pp.163–168.

Waldhauser, C., Hochreiter, R., Otepka, J., Pfeifer, N., Ghuffar, S., & Korzeniowska, K., et al., 2014. Automated Classification of Airborne Laser Scanning Point Clouds. *Solving Computationally Expensive Engineering Problems*. Springer International Publishing.

Zaletnyik, P., Laky, S., Toth, C., 2010. Lidar waveform classification using Self-organizing map. *ASPRS 2010 Annual Conference*.

Zhou, M., Li, C. R., Ma, L., & Guan, H. C., 2016. Land cover classification from full-waveform lidar data based on support vector machines. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B3*, pp.447-452.

Zhou, M., Liu, M., Zhang, Z., Ma, L., & Zhang, H.,2015. The research of land covers classification based on waveform features correction of full-waveform LiDAR. *SPIE Remote Sensing*, pp.96431P.