International Journal of Engineering and Geosciences (IJEG),
Vol;4, Issue;1, pp. 045-051, February, 2019, ISSN 2548-0960, Turkey,
DOI: 10.26833/ijeg.440828

AIRBORNE LIDAR DATA CLASSIFICATION IN COMPLEX URBAN AREA USING RANDOM FOREST: A CASE STUDY OF BERGAMA, TURKEY

Sibel Canaz Sevgen [1]*

[1]Ankara University, Faculty of Applied Sciences, Department of Real Estate Development and Management, Ankara, Turkey (ssevgen@ankara.edu.tr); **ORCID 0000-0001-5552-6067**

*Corresponding Author, Received: 05/07/2018, Accepted: 06/08/2018

ABSTRACT: Airborne Light Detection and Ranging (LiDAR) data have been increasingly used for classification of urban areas in the last decades. Classification of urban areas is especially crucial to separate the area into classes for urban planning, mapping, and change detection monitoring purposes. In this study, an airborne LiDAR data of a complex urban area from Bergama District, İzmir, Turkey were classified into four classes; buildings, trees, asphalt road, and ground. Random Forest (RF) supervised classification method is selected as classification algorithm and pixel-wise classification was performed. Ground truth of the area was generated by digitizing classes into features to select training data and to validate the results. The selected study area from Bergama district is complex in urban planning of buildings, road, and ground. The buildings are very close to each other, and trees are also very close to buildings and sometimes cover the rooftops of buildings. The most challenging part of this study is to generate ground truth in such a complex area. According to the obtained classification results, the overall accuracy of the results is found as 70,20%. The experimental results showed that the algorithm promises reliable results to classify airborne LiDAR data into classes in a complex urban area.

Keywords: *Random Forest, LiDAR, Classification, Complex Urban Area*

1. INTRODUCTION

Classification of objects in an urban area is a popular subject in a variety of research areas, such as computer vision, machine learning, pattern recognition, photogrammetry, remote sensing, and urban planning. In the literature, satellite and aerial images have been widely used in urban area classification. Especially, land cover changes studies over the years by classifying satellite data are abundantly presence in the literature. For instance, Yu et al. (2012) monitored land cover changes and urban sprawl dynamics 1989, 1999, and 2009 of Yantai China by classifying satellite images in five classes. Atlanta, Georgia's land cover changes 1973-1998 were categorized into six different classes (Yang and Lo, 2002). Canaz et al. (2017), classified Istanbul, Turkey, in four different classes to monitor land cover change between the years of 1986-2015. On the other hand, comparing with optical sensor data, a new technology to collect remotely sensed data is called as Light Detection and Ranging (LiDAR) have also been subjected as a popular data for classification studies. LiDAR technology is capable of collecting 3 Dimensional (3D) point cloud data in a short time day or night. Because of the direct 3D data acquisition, LiDAR data also have been increasingly used for classifying urban areas into classes.

Classical data-driven techniques have been developed for urban area classification (Rottensteiner and Briese, 2002, Charaniya et al. 2004), the recent trend is to use machine learning techniques to classify LiDAR data in urban area (Lodha et al., 2006). Supervised machine learning techniques are based on selected features and classifier algorithm. In the literature, a variety of supervised classification techniques, support vector machines, neural networks, exists (Richards, J.A., and Jia), in this study one of the supervised classification technique, called as Random Forest (RF) was selected and used because of its stability and robustness to the features.

RF classification for airborne LiDAR data has been studied using different features in order to label different classes. For instance, Niemeyer et al. (2012) classified three different area from Vaihingen, Germany LiDAR dataset named as 'ISPRS Test Project on Urban Classification and 3D Building Reconstruction'. The authors classified data into five categories; building, low vegetation, tree, terrain, and asphalt ground using Conditional Random Field (CRF) approach. However, they only showed and evaluated the result only for classes building and tree. Their correction result for classification for the 3 subset area of the data was found in average 73% and 92% for the tree and building classes, respectively. Guo et al. (2010) use a combination of optical multispectral and LiDAR data to classify LiDAR data in urban area in four classes using the Random Forest (RF) algorithm. Many other studies using RF algorithm to classify LiDAR data can be found in the literature (Immitzer, et al., 2012; Rodriguez-Galiano et al. 2012; Guan et al., 2013).

Lodha et al. (2006) employed another LiDAR data classification work. The authors used Support Vector Machines (SVM) for classifying LiDAR data into buildings, trees, roads, and grass using five features: height, height variation, normal variation, LiDAR return intensity, and image intensity. To evaluate result they compare ground truth and the classification result and observed 90% accuracy. Chen et al. (2013) classified LiDAR data to detect landslides in Three Gorges, China by using the mean aspect, Digital Terrain Model (DTM), and slope textures based on four texture directions; aspect, DTM, and slope textures based on aspect; and the moving average and standard deviation (stdev) filter of aspect, DTM, and slope and RF algorithm. By combine feature selection method with RF algorithm, they found a reliable result for classifying LiDAR data and detection of landslides. Ma et al. (2017) studied a comparison between SVM and RF algorithm to classify LiDAR data. The authors classified data in four categories: trees, buildings, farmland, and ground. According to their findings, the RF algorithm gave a better result than the SVM algorithm for the classification of the LiDAR data.

In this study, an area from the Bergama district of İzmir province, Turkey was chosen as study area. The study area is very complex in shape. The feature classes in interest are located very close to each other and some buildings and trees are embedded. Thus, the originality of the study is that the selected study area is very complex in shape. Therefore, digitization and generation of ground truth for the study area were carried out very carefully. After creating the ground truth and 12 features (which were generated from LiDAR data such as intensity, planarity, DSM etc.) were used to employ classification of LiDAR data.

2. STUDY AREA AND DATA

The study area was chosen from Bergama District of İzmir. İzmir is one of the biggest provinces in Turkey and located in western Turkey. Bergama is the biggest district of İzmir in the size of the area. The area of Bergama is 1573 km^2. The population of the district in 2017 is 102.961.

The study area is located in the center of Bergama district (Fig. 1). The boundary of İzmir province is shown with the blue line, and the boundary of Bergama district is shown in red line in Fig. 1. The true orthophoto of the study area is also shown in Fig. 1. Since the study area's land cover mainly consists of ground, roads, trees, and buildings, the study area divided four groups for classification: buildings, trees, ground, and asphalt road.

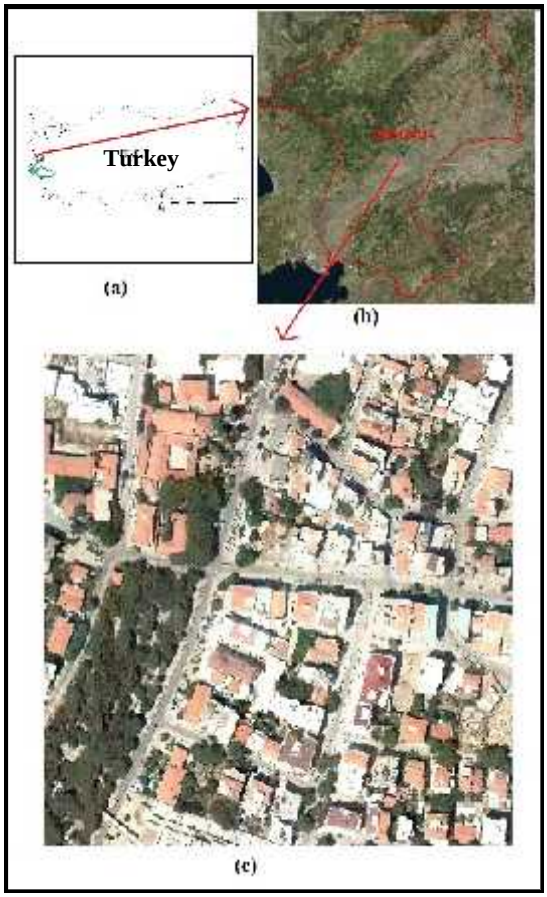


Figure 1. Location of study area, İzmir province boundaries (a), Bergama district boundaries (b) *(source: google maps)*, and the chosen study area (c)

True orthophotos of the study area were generated by Directory of Geographic Information Systems. The images were acquired in May 2016. The pixel size of the images is 10 cm. LiDAR data of the study area was collected by Optech Pegasus HA-500 technic by Turkish General Command of Mapping on 20-21 October 2014 (Kayı et al. 2015). Detailed information about the Optech Pegasus HA-500 is given in Table 1 (Optech, 2018).

Table 1. Technical information of Optech Pegasus HA-500 (Kayı et al. 2015)

Feature	Value
Height	150-5000 m
Effective laser repetition rate	100-500kHz
Scanning Angle	0-75º Adjustable
Accuracy (KOH)	≤ 5-20 cm.
Scanning Mechanism	Oscillating

3. METHODOLOGY

RF algorithm is often used in remote sensing applications to classify data such as multi and hyperspectral images, radar, LiDAR and thermal data sets. A literature review of these applications was presented in Belgiu and Dragut article (2016). This study is based on RF on one of the remote sensing data airborne LiDAR for a complex urban area. The flowchart of the methodology of this study is given in Figure 2.

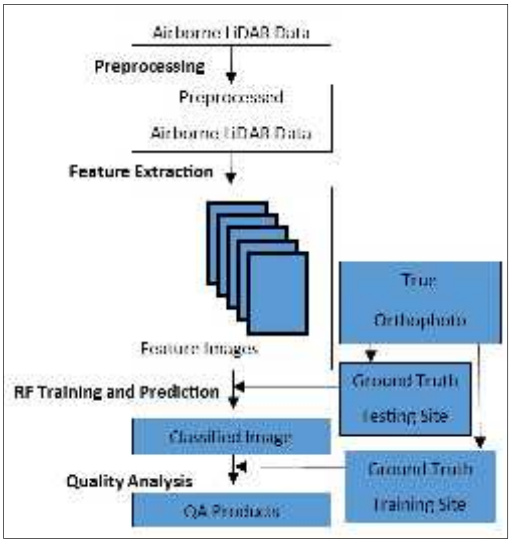


Figure 2. Flowchart of the methodology

The classification of the LiDAR data involves pixel-based classification; therefore, 12 features were generated and rasterized to 50 cm images. Before generating features images and classifying the study area, airborne LiDAR data was cleaned from noisy and duplicate points. After preprocessing, feature images were generated in four groups (Chehata et al., 2009; Dittrich et al., 2017) intensity, height, eigenvalue, and echo based. The intensity-based feature relies on the reflected energy of the objects in the LiDAR dataset. It helps to separate different characteristics objects such as asphalt road and ground classes. Intensity feature image is created using ArcGIS "Las to Raster" tool in 50 cm pixel size. Height based features, on the other hand, were generated from height values of the points and they play a really important role in separating ground and other non-ground classes, such as buildings and trees. The lidar data set was filtered to ground points, then from those points, a DTM in 0.5 m pixel size was generated. In addition to that, a 50 cm Digital Surface Model (DSM) was generated from all points in the LiDAR dataset. Normalized DSM (nDSM) was obtained by subtracting DSM from DTM. Besides, height features based on local neighborhood helps to determine objects, which are also different levels of the surface. *minh*, minimum height value in the neighborhood, and *Hd*, height difference from *minh* of the interested point, were generated for each point in the LiDAR dataset (Table 2). From those features, 0.5 m feature raster were generated using Python programming language (Python, version 2.7).

Table 2. Height based features

Feature	**Description**
nDSM	Normalized Digital Surface Model
minh	Minimum height in local neighborhood of a point
hd	The difference between minimum height in the local neighborhood of a point and that point height

Eigen-value based features were obtained from eigenvalues which were calculated from the local neighborhood covariance matrix. Eigen-values describe the shape of the object, thus they give valuable information about the object, whether it is a plane, line or sphere; therefore, those features are a good indicator of a tree or building roofs, depending on the feature (Table 3). Sphericity, *S*, planarity, *P*, linearity, *L*, anisotropy, *A*, the sum of eigenvalues, *Sum*, and change of curvature, *C*, were calculated and 0.5 m feature images for each feature were generated using Python Programming Language. Geometric features, sphericity, planarity, linearity, and anisotropy describe the shape of the object and give useful information about the object whether it is a line, plane or sphere. All geometric features were created in 3 m neighborhood points per point and then rasterized into 1 m range of mean values.

Table 3. Eigen-value based features

Feature	Description
Anisotropy	$\frac{\lambda 1 - \lambda 3}{\lambda 1}$
Planarity	$\frac{\lambda 2 - \lambda 1}{\lambda 1}$
Sphericity	$\frac{\lambda 3}{\lambda 1}$
Linearity	$\frac{\lambda 1 - \lambda 2}{\lambda 1}$
Change of curvature	$\frac{\lambda_3}{\sum_{i=1}^{3} \lambda_i}$
Sum of eigenvalues	$\sum_{i=1}^{3} \lambda_i$

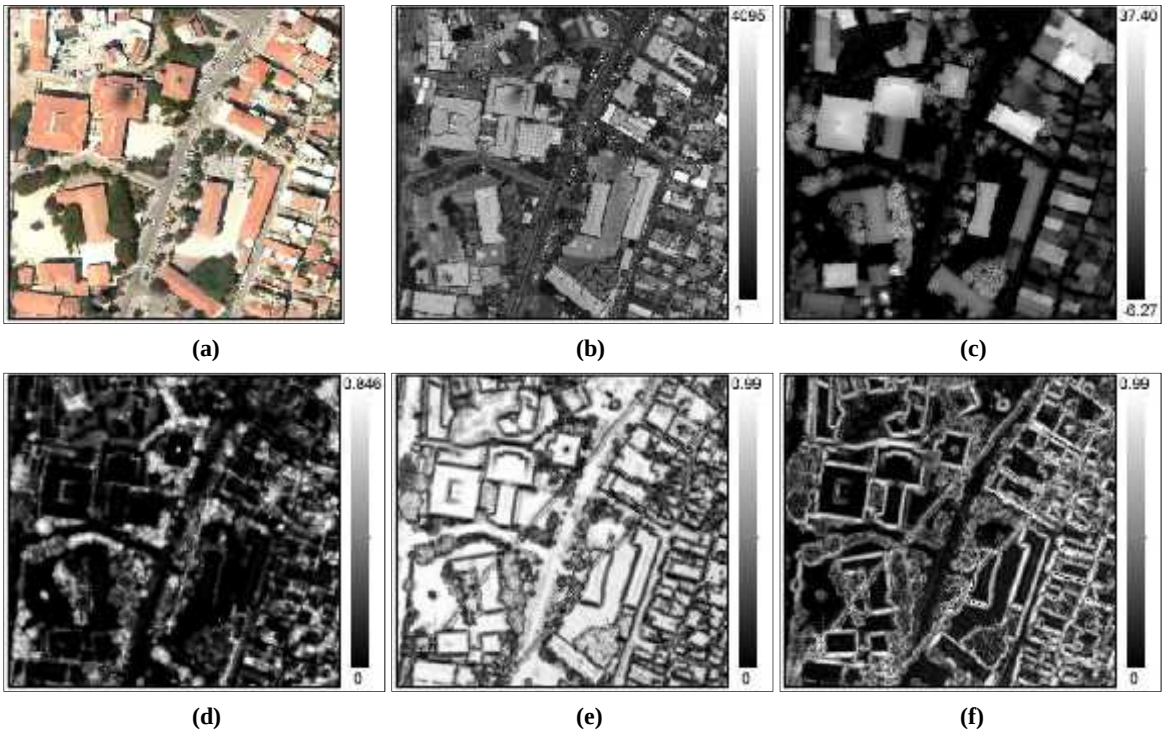
Last feature set, echo based features, helps to differentiate objects, which have multiple returns. Therefore, a total number of return, *n*, and the ratio of a number of return over a total number of returns, *t/n*, were calculated for each point and rasterized to 0.5 m images (Table 4).

Table 4. Echo based features

Feature	Description
n	Total number of returns
t /n	Number of returns over a total number of returns

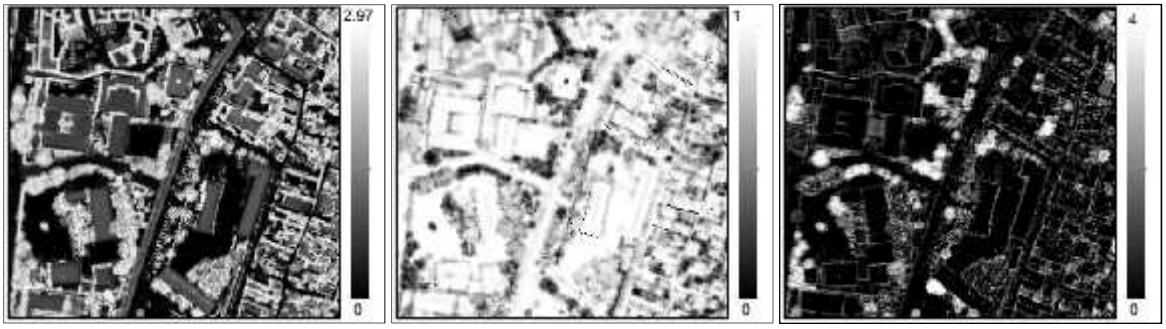
A total number of twelve features was selected and images were generated using Python programming Language and its machine learning and geospatial libraries, including scikit learn (Pedregosa et al., 2011) and GDAL (GDAL, 2018). Some of the features and orthophoto of a part of the study area are shown in Figure 3.

RF classification (Breimen, 2001) is an ensemble method of decision trees, which relies on randomly selecting a subset of features and creating multiple trees in training. and predicting new unlabeled data by voting each tree in the ensemble. Two parameters are required by the user, a number of trees, that define how much a tree can grow up and number of features, which determine how many new nodes can be split from parent node in the tree.



(a) **(b)** **(c)**

(d) **(e)** **(f)**



(g) **(h)** **(i)**

Figure 3. (a) True orthophoto and example of generated feature images; (b) intensity, (c) nDSM, (d) sphericity, (e) planarity, (f) linearity, (g) total number of returns, (h) anisotropy, (i) number of total returns over number of return images.

Ground truth of the study area (red boundary) and the training area (blue boundary) are shown in Figure 4. The study area and training areas were chosen from a different area. According to the similar studies in the literature, the size of the training area was chosen as no lower than the following size: *0,3 x size of the study area*. The study area was fully digitized to use it for quality control of the classification results. Pink, green, black, and yellow colored features represent buildings, trees, asphalt road, and ground, respectively.

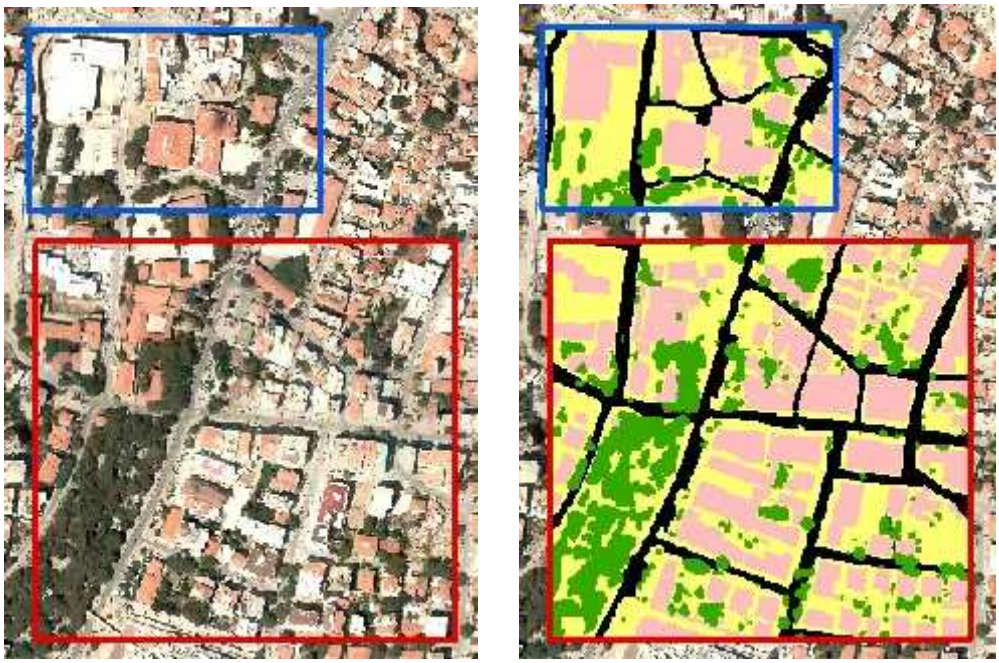


Figure 4. Study area (red), training area (blue) and their manually digitized features

Using the manually digitized training area (blue boundary Fig. 4) and the twelve features, the classification results were acquired by the RF algorithm. The results are described in the following section.

4. RESULTS

Ground truth of the area was created by digitizing the features from orthophoto of the area. Buildings, trees, asphalt road, and the ground were carefully digitized (Figure 5a). A part of the ground truth is used for classification as a training site, while the ground truth of the study area is used for the quality control of the results.

The results were classified into four groups is shown in Figure 5b. In the figure, red, green, gray and blue represent the buildings, trees, asphalt road, and ground classification results, respectively. As it can be seen in figure 5, the classes are extracted with high accuracy by comparing the proposed methodology classification results and the orthophoto of the study area. The qualitative analysis was employed by comparing the ground truth and the classification result. For this

purpose, the difference between the ground truth and result of the classes were created and illustrated in the Fig. 5c.

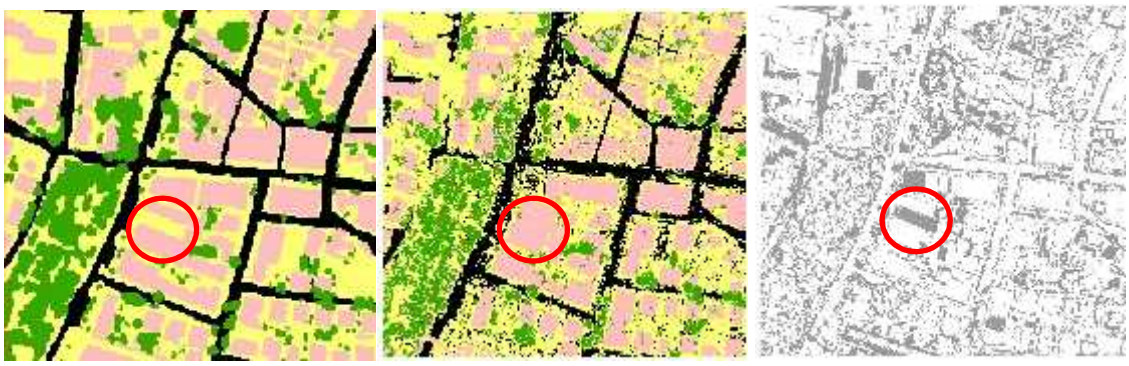


Figure 5. (a) Ground truth, (b) Classification result, (c) Difference

By using the ground truth and the classification results, the quality control was employed, and a confusion matrix was calculated. According to the results, the accuracy is found as, 77,90% , 58,37% , 72,90%, and 71,53% for the buildings, trees, asphalt road, and the ground, respectively. Overall accuracy for the results is 70,20%. Although the area is very complex, the classification results are reliable. Only trees class have lower results than results of the other classes. Some errors occurred since LiDAR data, and the orthophotos, which were used to create ground truth, were acquired in different years and seasons. Therefore, some of the trees might misclassified just because in LiDAR data acquisition time (October 2014), the trees might not have leaves on the trees. On the other hand, orthophotos were collected and created in May, when trees have leaves. Another reason that affects the results is that, in orthophotos, some of the buildings were demolished while they are present in the LiDAR data. For instance, one of the case for this kind of building is shown in Figure 5 with red circles. Finally, there were cars that were on the roads in the LiDAR data, while they are not presence in the orthophoto. This phoneme also mismatch the classification of asphalt roads

5. CONCLUSION

In conclusion, in this study, LiDAR data of a complex urban area from Bergama district, İzmir, Turkey was classified into four groups using the RF algorithm. The classes are as following, buildings, trees, asphalt road, and ground. The area is very complex in terms of city planning for instance buildings' shapes are irregular. The most challenging part for this study was a generation of the ground truth since the area is very complex in shape. Digitization of roads and buildings was very difficult and carried out very carefully. After digitization of the area, twelve features were created from LiDAR data, and using the features and ground truth together, the area is classified by RF algorithm. According to the results, the RF algorithm was classified the area reliably with 70,20% overall accuracy. However, some errors occur because the LiDAR data was acquired in October 2014 and the orthophoto used in this study was collected in May 2016. Because of the seasonal effect, some of the trees were not classified by the proposed methodology. Moreover, in some cases, some building and trees that are available in the orthophoto images, is not found in the LiDAR data, Finally, for the asphalt road, there are car on the roads, which may not be on the LiDAR data or vice versa. These affected the classification results. Even though, these limitations, the proposed methodology is able to classify the complex urban area with high accuracy.

ACKNOWLEDGEMENTS

The author is very thankful to Eray Sevgen, a Ph.D. student at the Hacettepe University for sharing Python scripts for the RF algorithm, helping the feature extraction and digitizing of ground truth data. The author is also thankful to the Turkish Directory of Geographic Information Systems for providing true orthophoto images of the study area and the Turkish General Command of Mapping for providing the LiDAR data of Bergama district.

REFERENCES

Belgiu, M. and Drăguț L. (2016). Random forests in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens., 114 , pp. 24-31.

Breiman, L. (2001). "Random Forests." Machine Learning 45: 5–32.

Canaz S., Aliefendioğlu Y. and Tanrıvermiş H. (2017). Change detection using Landsat images and an analysis of the linkages between the change and property tax values in the Istanbul Province of Turkey. Journal of Environmental Management. Vol. 200:446-45.

Chehata, N., Li, G. and Mallet, C. (2009). Airborne LIDAR feature selection for urban classification using random forests. Geomat. Inform. Sci. Wuhan Univ. 38, 207–212.

Dittrich, A., Weinmann, M. and Hinz, S. (2017). Analytical and numerical investigations on the accuracy and robustness of geometric features extracted from 3D point cloud data. ISPRS J. Photogramm. 126, 195–208.

Charaniya, A.P., Manduchi, R. and Lodha, S.K. (2004). Supervised parametric classification of aerial LiDAR data. In Proceedings of 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), Washington, DC.

Chen, W., Li, X., Wang, Y., Chen, G. and Liu S. (2014). Forested landslide detection using LiDAR data and the random forest algorithm: a case study of the Three Gorges, China. Remote Sens. Environ. 152 (2014), pp. 291-301.

GDAL/OGR contributors (2018). GDAL/OGR Geospatial Data Abstraction software Library. The Open Source Geospatial Foundation. URL http://gdal.org

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z. and Yang, X. (2013). Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. International Journal of Remote Sensing, vol. 34, issue 14, pp. 5166-5186.

Guo, L., Chehata, N., Mallet, C. and Boukir, S. (2011). Relevance of airborne LiDAR and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing, 66 (1) , pp. 56-66/

Kayı A. Erdoğan M. and Eker O. (2015). Results of LİDAR test performed by OPTECH HA-500 and RIEGL LMS-Q1560. Harita Dergisi, Volume 153, pp 42-46.

Lodha, S.K., Kreps, E.J., Helmbold, D.P. and Fitzpatrick, D. (2006). Aerial LiDAR data classification using support vector machines (SVM). The Third International Symposium on 3D Data Processing, Visualization, and Transmission pp. 567-574.

Ma L., Zhou M. and Li C. (2017). Land Covers Classification Based On Random Forest Method Using Features From Full-Waveform Lidar Data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W7, 2017 ISPRS Geospatial Week 2017, 18–22 September 2017, Wuhan, China

Niemeyer, J., Rottensteiner, F. and Soergel, U. (2012). Conditional random fields for LiDAR point cloud classification in complex urban areas. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, pp. 263-26.

Optech, (2018), http://www.Optech.Com/Specification [Accessed on: 14 May 2018]

Richards, J.A. and Jia, X. (1999). Supervised Classification Techniques Remote Sensing Digital Image, Analysis, Springer-Verlag GmbH, Heidelberg (1999) pp. 193–247.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blonde,l M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:2825-2830.

Python Software Foundation. Python Language Reference, version 2.7. available at http://www.python.org. Retrieved on 15.04.2018.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67 (0), pp. 93-104.

Rottensteiner, F. and Briese C. (2002). A new method for building extraction in urban areas from high-resolution LiDAR data. Int Arch Photogramm Remote Sens Spat Inf Sci 34(3A):295–301

Yang, X. and Lo, C.P. (2002). Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area , International Journal of Remote Sensing, 23, pp. 1775—1798.

Yu, X., Zhang, A., Hou, X., Li, M., and Xia, Y. (2013). Multi-temporal remote sensing of land cover change and urban sprawl in the coastal city of Yantai, China. International Journal of Digital Earth. Vol. 6, Supplement 2, 137-154.