CLASSIFICATION OF DATA FROM AIRBORNE LIDAR BATHYMETRY WITH RANDOM FOREST ALGORITHM BASED ON DIFFERENT FEATURE VECTORS

T. Kogut [1], M. Weistock [1], K. Bakuła[2]

[1] Dept. of Geoinformatics, Koszalin University of Technology, Sniadeckich 2, 75-453 Koszalin, Poland - tomasz.kogut@tu.koszalin.pl, marlena.weistock@gmail.com
[2] Warsaw University of Technology, Faculty of Geodesy and Cartography, Department of Photogrammetry, Remote Sensing and Spatial Information Systems, 00-661 Warsaw, Poland - krzysztof.bakula@pw.edu.pl

KEY WORDS: classification, random forest, airborne lidar bathymetry, object detection, full waveform

ABSTRACT:

Modern full-waveform laser bathymetric scanners offer the possibility of a practical application of airborne laser bathymetry (ALB) data algorithms as a valuable source of information in the study of the aquatic environment. The reliability of the obtained results and the efficiency of the classification depend on the applied features. The input data for the classifier should consist of variables that have the ability to discriminate within the data set, for the detection and classification of objects on the seabed. The automatic detection of underwater objects is based on machine learning solutions. In this paper, the ALB data were used to present a classification process based on the random forest algorithm. The classification was carried out using two independent approaches with two feature vectors. The quality of classifications based on the full-waveform features vector and the geometric features vector was compared. The efficiency of each classification was verified using a confusion matrix. The obtained efficiency of the point classification in both cases was about 100% for the water surface, 99.9% for the seabed and about 60% for underwater objects. Better results for the classification of objects were obtained for the features vector based on features obtained directly from full-waveform data than for the vector obtained from geometric relationships in the point cloud.

1. INTRODUCTION

Airborne laser bathymetry (ALB) is a technique which has become more and more important in recent years, as a result of improved hardware and better processing software. This measurement method can be used as a supplement or a complete alternative to sonar data (Steinbacher et al., 2010). Bathymetric lidar is a multispectral lidar system (Bakuła, 2015), equipped with a green laser beam with a wavelength of 532 nm or a combination of a green and an infrared beam with a wavelength of 1,064 nm, measuring the full waveform. The use of an additional infrared beam allows the derivation of some information about the water surface, because the beam is not able to penetrate the water column and sufficient energy for its return is not guaranteed (Mandlburger et al., 2013). Classification is a typical problem for machine learning approaches, which are constantly developing and gaining in popularity due to the large number of tasks they can perform. Successful, automatic data classification can significantly contribute to the development of effective methods for monitoring obstacles on the sea bottom. Knowledge about the structure of the bottom in coastal areas can affect their continued functioning. Each classification is based on a set of features provided to the classifier, which performs the decision-making process and returns the result. In practice, it is not always known which of these features are important for the classification. Therefore, a wide range of variables is usually chosen, to increase the efficiency of the classifier. However, a large number of features does not always ensure that the identification of the class is perfect. The correct choice of features significantly influences the quality of the classification (Hastie, Tibshirani and Friedman, 2009). The IHO standard ISO S-44 imposes the obligation of detecting objects on the seabed depending on the depth and purpose of the investigated body of water. With this in mind, the point cloud was classified into three classes: the water surface, the bottom and an object on the sea bottom. The goal of this article is to examine two feature vectors that enable effective classification of airborne laser bathymetry data and hence detection of objects on the seabed. This paper presents the classification of the point cloud using the random forest algorithm based on two feature vectors. Based on an analysis of the research presented by Chehata et al. (2009), Gross et al. (2006) and Niemeyer et al. (2014), the features applicable in topographic lidar were selected. Two groups of features can be distinguished. The first is based on the geometry of the point cloud and the second is based on the full-waveform data. This study has been carried out using data collected as part of the project “Investigation on the Use of Airborne Laser Bathymetry in Hydrographic Surveying”, which was undertaken by the Federal Maritime and Hydrographic Agency of Germany in Rostock and the Institute of Photogrammetry and GeoInformation at Leibniz University in Hannover.

2. METHODOLOGY

The following section contains a description of the random forest algorithm used in the experiment. The feature vectors are defined and explained in the next subsection. Finally, the test area is presented.

2.1. Random Forest Algorithm

The classification of the point cloud was based on the features that can be obtained from the point cloud geometry or directly from the full waveform. The classification was carried out for the two independent approaches in MATLAB, using the random forest algorithm (https://davidlary.info) with two feature

vectors. Points were assigned to one of three classes: water surface, seabed or object. The first approach was the classification of all points and the second was the classification of points excluding the water surface, which was filtered out on the basis of the value of $\Delta z > 1$ m (height difference). In the research carried out by Chehata et al. (2009), Guan et al. (2012) and Gan et al. (2015), the usefulness of the random forest algorithm for the classification of airborne laser scanning data covering urban areas was confirmed. Random forest is an advanced implementation of the bagging algorithm, which uses decision trees as the base model (Breiman, 2001). The concept of this algorithm is based on the construction of a group of decision trees, created on the basis of a random set of data from the training set. Each tree is trained on a bootstrap sample, and the attribute used to divide this sample is chosen from among a selected subset of variables. Their number is much smaller than the number of all features in the sample, and therefore the random forest algorithm can be used to classify sets with an observation vector of large dimension. The observation vector is classified by all trees and assigned to the class with the highest number of trees assigned to it. In the classification process, the parameter that is to be determined is the number of random variables m. Studies (Breiman, 2001, Chehata et al., 2009, Gislason et al., 2006) show that the value that allows good classification results to be obtained is the square root of all the features of the observation vector. An assessment of the probability of incorrect classification is built into the algorithm generation process. The classification error is estimated on the basis of training elements not included in the given bootstrap test set on which the classifier is trained. This sample consists of about two thirds of the original training set. Due to observations outside this pseudo-test (out-of-bag estimate), the random forest algorithm also indirectly evaluates the usefulness of features in the learning process (Liaw and Wiener, 2002). The random forest algorithm is characterized by stability, resistance to overfitting, effective operation on large data sets and the ability to detect interactions between variables.

2.2. Feature Vectors

In the classification of airborne laser bathymetry data, features obtained from the analysis of the point cloud geometry and the full waveform were used. These features are shown in Table 1.

	Symbol	Feature
Full-waveform features	A	amplitude
	ω	echo width
	N	return number
	N_t	number of returns
	N_z	normalized echo
Geometric features	Δ_z	height difference
	σ_z^2	height variance
	$\lambda_1, \lambda_2, \lambda_3$	eigenvalues
	S_λ	sphericity
	P_λ	planarity
	L_λ	linearity
	A_λ	anisotropy
	E_λ	eigentropy
	$O_{v\lambda}$	omnivariance

Table 1. Features used in classification

The combination of features used in the analysed classification process was determined by the results of previous experiments involving the classification of the ALB point clouds using the random forest algorithm with one vector consisting of 16 features (Kogut et al., 2019). The results of these experiments allow the suitability of individual features for distinguishing classes, to be determined. In the presented case, the classification was carried out using features divided into two vectors:

a feature vector based on the full waveform, defined according to Equation 1

$$f_1 = [A, \omega, N, Nt, Nz]^T \quad (1)$$

and a feature vector based on the geometry of the point cloud, defined according to Equation 2.

$$f_2 = [\Delta z, \sigma_z^2, \lambda 1, \lambda 2, \lambda 3, P_\lambda, S_\lambda, L_\lambda, A_\lambda, E_\lambda, O_{v\lambda}]^T \quad (2)$$

The aim of the classification based on these vectors was to select a set of features that allow the efficiency of the classification of ALB data to be improved. Due to the specificity of ALB data with regard to their classification, it is necessary to select those features that represent a significant property of the studied area. In this investigation, the features based on the full waveform were determined by an approximation of the waveform to a Gaussian function (Wagner et al., 2006) and the geometric features of the point cloud were determined by testing the dependencies of points in the neighbourhood defined by a cylinder with a 5-m radius. Due to the specific character of the point cloud in airborne lidar bathymetry, where two major layers occur (points on the water surface and points on the seabed), some features (σ_z^2, λ_1, λ_2, λ_3, P_λ, S_λ, L_λ, A_λ, E_λ, $O_{v\lambda}$) have been determined only for points under the water surface. These were used in both classification attempts.

A description of the features, showing how they are potentially useful in the classification, is given below.

- *A*: Amplitude. This assumes higher values for reflections from the water surface.
- *ω*: Echo width. This feature is clearly higher for reflections from the seabed in ALB.
- *N*: Return number.
- N_t: Number of returns. In airborne lidar bathymetry, this feature potentially assumes high values for objects on the seabed.
- N_z: Normalized number of echoes, obtained by dividing the echo number by the total number of echoes in the full waveform of the current point.

Δ_z: Depth difference between the tested point and the point located at the lowest height inside the cylinder. This feature is clearly higher for points on the water surface. On the bottom, the calculated difference is small.

- σ_z^2: Height variance.
- λ_1, λ_2, λ_3: Eigenvalues. The point cloud was enriched with additional information through determination of the following features (Gross et al., 2006).
- P_λ: Planarity. This determines planes in the data set and assumes different values for objects protruding above the seabed. It is defined by Equation 3.

$$P_\lambda = \frac{\lambda_2 - \lambda_3}{\lambda_1} \quad (3)$$

- S_λ: Sphericity. This is a feature explaining spatial elements in the data set, defined by Equation 4.

$$S_\lambda = \frac{\lambda_3}{\lambda_1} \tag{4}$$

- L_λ: Linearity, indicating the linear nature of the point distribution of objects located on the seabed, defined by Equation 5.

$$L_\lambda = \frac{\lambda_1 - \lambda_2}{\lambda_1} \tag{5}$$

- A_λ: Anisotropy, specifying the relationships between the directions of the point distribution, defined by Equation 6.

$$A_\lambda = \frac{\lambda_1 - \lambda_3}{\lambda_1} \tag{6}$$

- E_λ: Eigentropy, i.e., the entropy calculated from the eigenvalues defined by Equation 7.

$$E_\lambda = -\sum_{i=1}^{3} \lambda_i \ln \lambda_i \tag{7}$$

- $O_{v\lambda}$: Omnivariance. High values of this feature correspond to spatial objects and low values to plane areas or linear structures, defined according to Equation 8.

$$O_\lambda = \sqrt[3]{\prod_{i=1}^{3} \lambda_i} \tag{8}$$

2.3. Test Field

The presented research was carried out using data covering the area of the Rosenort artificial reef located in the Baltic Sea, about 25 km north of the city of Rostock (Germany). The Rosenort reef is made up of four artificially created zones. The individual reef zones are formed by 50 two-ton concrete tetrapods, 180 tons of stones, 30 cut-concrete cones and six six-ton concrete tetrapods. The reef is located about 2,000 m offshore.

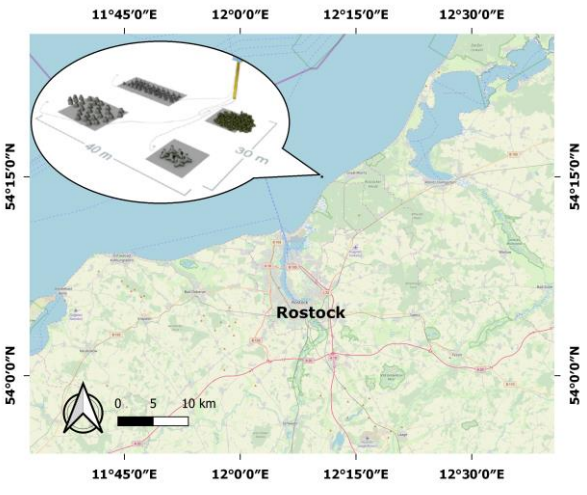

Figure 1. Location of the test object, Baltic Sea.

The data were obtained in August 2013, using the AHAB Chiroptera scanner. The mean density of points for the object is three points per m^2. A total of 30,594 points were measured with a mean depth of about 6 m. Reference data for the analysis of conformity of results were obtained via manually performed classification of the data set.

3. RESULTS OF CLASSIFICATION

The results of the random forest classification were used to compare the suitability of the two feature vectors for the classification of ALB data. For each vector, the classification was carried out twice. In the first approach, all points in the point cloud were classified and in the second approach only the points not on the water surface were classified. In both cases, artificially created objects on the reef were used as the training area. After the classification, an analysis of its correctness was performed. The quantitative classification efficiency for the investigated vectors was analysed on the basis of the confusion matrix (Tables 2-5).

	Water surface [point]	Seabed [point]	Object [point]	**Accuracy [%]**
Water surface	10576	36	0	**99,6**
Seabed	0	13305	13	**99,9**
Object	0	82	130	**61,3**

Table 1. Confusion matrix and the result of classification of all points with full-waveform features

	Water surface [points]	Seabed [points]	Object [points]	**Accuracy [%]**
Water surface	10612	0	0	**100**
Seabed	0	13303	15	**99,9**
Object	0	93	119	**56,1**

Table 2. Confusion matrix and the result of classification of all points with geometric features

	Seabed [points]	Object [points]	**Accuracy [%]**
Seabed	13303	15	99,9
Object	78	134	63,2

Table 3. Confusion matrix and the result of classification of points excluding the water surface with full-waveform features

	Seabed [points]	Object [points]	**Accuracy [%]**
Seabed	13306	12	99,9

Object	86	126	59,4

Table 4. Confusion matrix and the result of classification of points excluding the water surface with geometric features

The results show that the random forest algorithm achieved a similar accuracy for each combination of features (each vector). In both variants of the feature vector, the accuracy of points classification on the bottom was 99.9%. Poorer results for the water surface classification (99.6%) were obtained from the data trained with the full-waveform features vector, which translated into an incorrect classification of 36 out of 10,612 points. The points on the objects were identified with the lowest accuracy, resulting in a classification effectiveness in all cases of about 60%. For the classification of all points with the geometric features vector, an incorrect classification occurred for 93 out of 212 points. This represented the poorest result of 56.1% for the classification of points on the objects. Using the features based on the full waveform increased the accuracy of classification for the class of objects to 61.3% (in the classification of the whole point cloud) and to 63.2% (in the classification of points excluding the water surface).

4. VALUE SIGNIFICANCE RANKING

By using the properties of the random forest algorithm, it is possible to determine the importance of the features used for its learning. The classifier selects features from the model in a stochastic manner, and therefore estimates the significance of less important variables (Guyon and Elisseeff, 2003). Using the algorithm before the construction of the tree from the original training data set, a set of features is drawn from the return. If there are no selected elements, this creates a sort of out of bugs. It is possible then to measure the significance of variables. The value significance ranking allows the features with the best discriminating properties to be selected. Rejecting the features that do not affect the results of the classification can improve the accuracy of the classification and reduce the complexity of the algorithm. The information on the feature significance is obtained directly from the trained classifier. The potential relevance of the value significance ranking can only be assessed in the classification process. The ordered features, depending on their significance in the classification of data from airborne laser bathymetry, are shown in Fig. 2 (for a vector of features based on the full waveform) and Fig. 3 (for a vector of features based on the geometry of the point cloud).

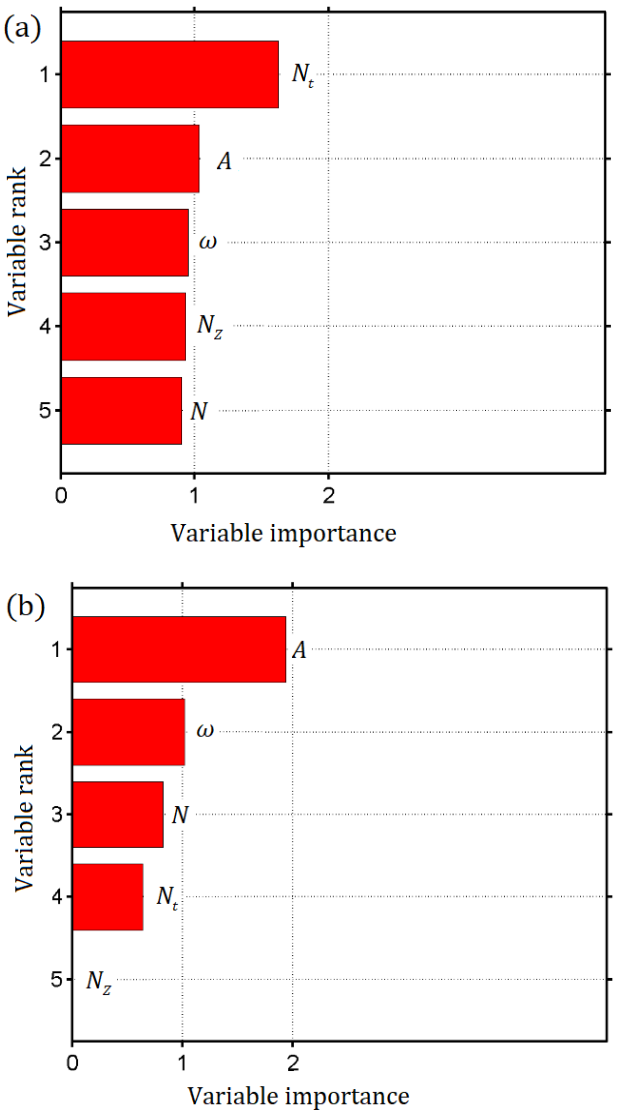

Figure 2. Vector features ranking f_1 for the classification of (a) all points and (b) points excluding water surface

Among the features based on the full waveform, the most important for the first classification variant is the number of returns. For the second variant, the most important feature is the amplitude. An important variable in both cases is the full width at half maximum. The number of returns is important for classification of all points in the point cloud, but in the classification of points excluding the water surface it is less important.

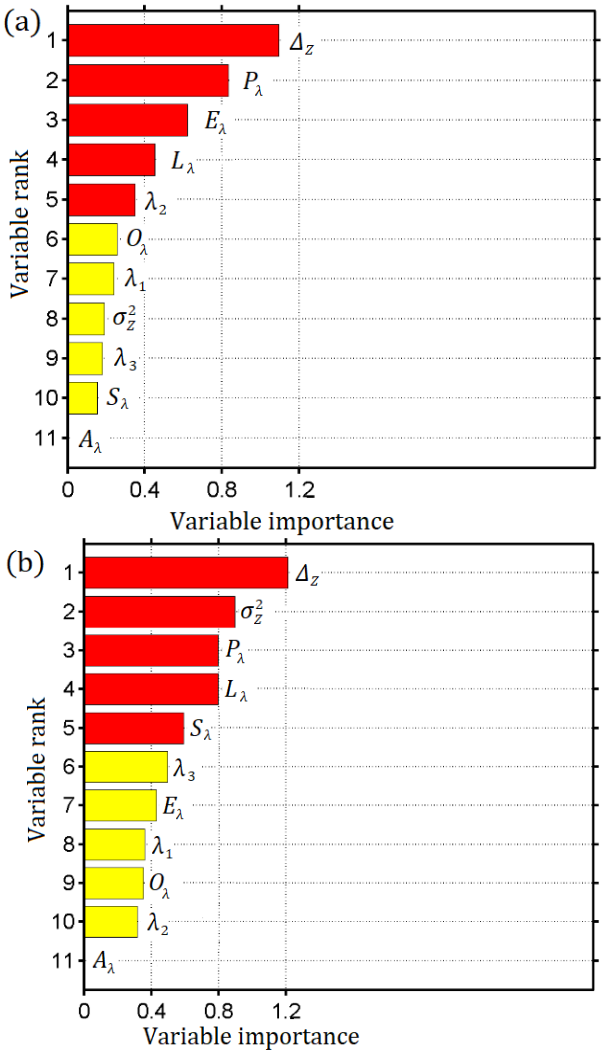

Figure 3. Vector features ranking f_2 for the classification of (a) all points and (b) points excluding water surface

Analysing above figures, it can be seen that the most important geometrical feature is the height difference. However, planarity is also important. The feature with lowest importance is the anisotropy.

5. CONCLUSION

In this paper, the random forest classifier was used to compare the suitability of two feature vectors for the problem of automatic detection of objects on the seabed. The proper selection of features for ALB data can create an effective tool for their classification. In this study, the effectiveness of the random forest algorithm for the vector of full-waveform features and the vector of features based on the geometry of the point cloud, was tested. In terms of classification effectiveness, the vector of features based on the full waveform was better. These features provide better results for the detection of objects located on the seabed. However, for the water surface and seabed points, the difference in accuracy is negligible. For the object class, an accuracy of 61.3% (compared with the reference results from manually classified points) was achieved. This value increased to 63.2% when the water surface points were excluded. The result obtained for this class using the random forest classification based on the vector of geometrical features, were lower, at 56.1% and 59.4% respectively. Based on these classifications, it can be concluded that the classification result depends on the selected feature vector. The use of source data without decomposition, if available, is recommended.

ACKNOWLEDGEMENTS

The authors would like to thank the Institute of Photogrammetry and GeoInformation in Hannover and the Federal Maritime and Hydrographic Agency of Germany in Rostock for the data used in this paper. Funding for the project “Investigation on the Use of Airborne Laser Bathymetry in Hydrographic Surveying” is provided by the Federal Maritime and Hydrographic Agency of Germany, under project number 10019311. The support is gratefully acknowledged.

REFERENCES

Bakuła, K., 2015. Multispectral airborne laser scanning - A new trend in the development of LiDAR technology. *Archiwum Fotogrametrii Kartografii i Teledetekcji* , 27, 25–41.

Breiman, L., 2001, Random Forests, Statistics Department University of California, Berkeley, CA 94720.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3/W8, 207-212.

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

Gan, Z., Zhong, L., Li, Y., Guan, H., 2015. A random forest based method for urban object classification using lidar data and aerial imagery. *23rd International Conference on Geoinformatics*, 2015, 1-4.

Gross, H., Thoennessen, U., 2006. Extraction of lines from laser point clouds. In *ISPRS Conference Photogrammetric Image Analysis (PIA)*, 36(3A), 87–91.

Guan, H., Yu, J., Li, J., Luo, L., 2012. Random forests-based feature selection for land-use classification using lidar data and orthoimagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, 39(B7), 203–208.

Guyon, I, Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.

Hastie, T., Tibshirani, R., Friedman, J., 2003. The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer.

Kogut, T., Wiestock, M., 2019. Classifying airborne bathymetry data using the Random Forest algorithm. *Remote Sensing Letters*, 10 (9), 874-882.

Liaw A., Wiener M., 2002, Classification and Regression by random Forest, *R News*, 2(3), 18 – 22.

Mandlburger, G., Pfennigbauer, M., Pfeifer, N., 2013. Analyzing near water surface penetration in laser bathymetry—Acase study at the River Pielach. *ISPRS Ann. Photogramm.Remote Sens. Spat. Inf. Sci.*, II-5/W2, 175–180.

Niemeyer, J., Rottensteiner, F., Sörgel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 152-165

Steinbacher, F., Pfennigbauer, M., Aufleger, M., Airborne hydromapping area-wide surveying of shallow water areas. In: Dittrich, Andreas; Koll, Katinka; Aberle, Jochen;Geisenhainer, Peter (Hg.): River Flow 2010. Karlsruhe: Bundesanstalt für Wasserbau. S.River Flow 2010, 1709–1714. Available online: https://hdl.handle.net/20.500.11970/99833 (accessed on day 20.05.2019)

Wagner, W., Ullrich, A., Ducic, V., Melzer, T., Studnicka, N., 2006. Gaussian decomposition and calibration of anovel small-footprint full-waveform digitising airborne laser scanner. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60, 100–112.