Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 11th International Scientific and Practical Conference. Volume II, 78-83

Geometric Feature Selection of Building Shape for Urban Classification

Sergejs Kodors
Rezekne Academy of Technologies

Abstract. ***The proposed research is related with building detection in airborne laser scanning data. The result of geospatial surface segmentation provides a vector layer of unclassified shapes. Geometric features of shapes can be applied to classify urban objects and to detect buildings among them. The goal of this research is to select the appropriate geometric features considering their importance for building recognition. The feature selection is completed using random forest algorithm. The obtained list of features and their influence weights can be used to improve building recognition methods and to filter noise objects.***

Keywords: feature selection, LiDAR, remote sensing, urban classification.

I. INTRODUCTION

Airborne laser scanning is the modern technology of remote sensing to acquire 3D model of Earth surface using aircraft and laser altimeter. The recorded 3D model is a point cloud, which points are detected locations of scanned object surface, where a laser beam is reflected. However, the acquired data are not applicable for geospatial analysis until a semantic meaning is assigned to them that is doable by classification methods developed for LiDAR data.

The methods of urban classification can be divided into two groups:

- 3D methods, which group near points into 3D clusters, classify them (e.g. the voxel-based method [1]) and return classified point cloud;
- and 2D methods, which use the projection of point cloud converted into 2D grid, that is segmented and classified (e.g. saliency-based method [2]) providing a classified raster or vector layers of classes.

This research is related with 2D methods. Different LiDAR features have been analysed and compared in the scientific publication [3], but these features are useful to convert the point cloud into 2D projection. When the segmentation of 2D projection and the classification of the obtained segments are completed, the vector layer of search objects is prepared. Depending on the used methods, the different amount of noise objects can be obtained together with search objects, but the vector layer provides the secondary data for the noise removing tasks. Therefore, it is possible to construct the classification workflow (see Fig.1), where clear data are obtained through the sequence of different feature analysis removing specific noise objects step by step.

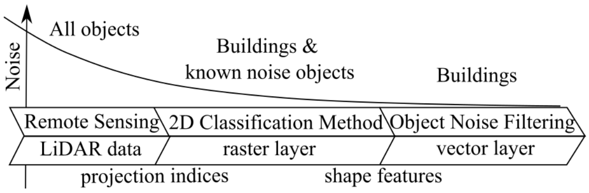


Fig. 1. Workflow of urban classification with three steps

There are three types of object features, which are applicable for secondary data classification using vector layer with shapes:

- geometric features;
- statistical features;
- spatial relations.

The goal of research is to analyse the geometric features of shape and to select the most appropriate of them for building detection and recognition tasks.

II. MATERIAL AND METHODS

A. Dataset

The LiDAR data of 25 km^2 territory have been used to complete research.

The minimal point density of samples is 1 p/m^2.

The coordinate system is LKS92 TM (EPSG:3059).

B. Dataset pre-processing

The dataset is processed in two stages:

1st stage: LiDAR data are segmented and classified by EMA-based method [2]. The output is the vector layer with buildings and noise objects. The known noise objects are bridges, huge engines, robust trees and shrubs, hedges, walls and cliffs.

2nd stage: All shapes are manually classified by two classes “building” or “noise object”.

ISSN 1691-5402

http://dx.doi.org/ 10.17770/etr2017vol2.2613

C. 2D classification method of 1st stage

The classification of LiDAR data is based on “Energy Minimization Approach” methodology [2].

The method applies the next algorithms:

- LiDAR data projection:

max-min method, which set pixel value equal to the height of single or last return point with the maximal height. Size of used grid for projection is 1 m^2.

- segmentation seeds:

the height difference points with the bias equal to 1.8 m, which is the most important feature according to the research [3].

- segmentation and classification:

4-path min-cut/max-flow Dinic’s algorithm, where objects are buildings and noise objects, background – ground.

- vectorization: 4-path Theo Pavlidis’ algorithm.

D. Data collections

The detected objects are manually verified and classified with label “building” or “noise” using the geographical information system called *Quantum GIS*. The total number of detected objects after the classification method is 844 284, where 99.68% are noise-objects (2658 objects are buildings only). The number of unique objects is 34 793 (only 4% from total number of objects), where 2484 are unique building shapes (7% from total number of unique samples). The most number of noise-objects are robust trees and shrubs (see Fig. 2).

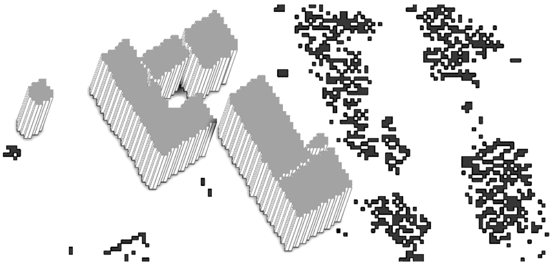


Fig. 2. Sample of building vector layer noised by robust vegetation elements

E. Analysed geometric features

Eleven geometric features are analysed under the scope of study, the list of them and their equations are provided in Table I. The geometric features are calculated using *Quantum GIS* and *OpenJump GIS* software, then the collected data are saved in *CSV* format to process them by data mining tools.

F. Feature importance analysis

The analysis are completed using *“R project”* tool. The correlation among features are analysed using Spearman monotonic correlation coefficient. The importance of features for object recognition is measured using random forest algorithm.

III. DISTRIBUTION ANALYSIS OF FEATURES

The distribution analysis of noise data show that data have two hills (*R*, *E, F*, C_2 and *S* cases). It means that two classes of data are grouped in one dataset. The feature eccentricity (*E3*) has the most strongly expressed distance between these subsets (see Fig.3).

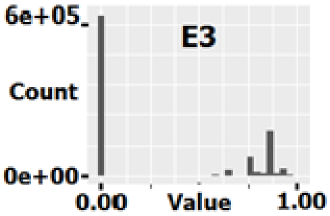


Fig. 3. Eccentricity distribution of noise samples, which shows existence of two subsets of noise classes

The anomaly subset has been selected by the logical expression: ($R > 0.875$) & ($E1 < 0.125$) & ($E2 > 1.05$) & ($E3 < 0.125$) & ($F > 0.875$) & ($C2 > 3.5$) & ($S > 0.95$). The obtained set contained 456 144 objects (54% of noise objects) and only one unique sample with area 1 m^2 (one pixel object).

IV. CORRELATION ANALYSIS OF FEATURES

Correlation among features are very important parameter in machine learning, because strongly correlated features don’t have additional information for classification and can be replaced by related other parameter with a goal to use the better minimal feature set for classification tasks, that sufficiently can minimize the processing time.

If the random forest is used for analysis of feature importance, the existence of correlation is very important factor, because the magnitude of feature importance is steadily decreasing, when the strongly correlated features are added to dataset [7].

Spearman correlation coefficients are calculated for unique object shapes. The buildings and noise objects are analysed independently (see Fig.3 and Fig.4).

The correlation analysis has showed, that building shapes have strong correlation among next features (see Fig.3): $\{A, P, C_2\}$, $\{C_1, K\}$, $\{E_1, E_2, E_3, F\}$ and $\{R, S\}$.

The noise objects have strong correlation among (see Fig.4): $\{A, P, C1, K\}$, $\{E_1, E_2, E_3, F\}$ and $\{R, S\}$.

The calculation complexity and intersection of value distribution are considered selecting the better feature of correlated sets (see Table I and II):

- $\{E_1, E_2, E_3, F\} \rightarrow F$ (form factor);
- $\{R, S\} \rightarrow R$ (rectangularity);
- $\{A, P, C_2\}$, $\{C_1, K\}$, $\{A, P, C1, K\} \rightarrow \{A$ (area), *C1* and *C2* (compactness)$\}$.

So, the suitable feature set is $\{A, R, F, C_1, C_2\}$.

V. RANDOM FOREST ALGORITHM

The classification task is to correctly assign class *y* to object *x*, where *X* is the set of objects and *Y* – the set of classes. So, the classification problem is to find the function, which most closely approximates function $f : X \to Y$.

Decision trees and random forest algorithms belong to supervised learning algorithms, when *m* samples are used to teach a classification system.

Decision trees are based on the graph theory. The goal of algorithms is to find the best rules of data classification. The result is the acyclic directed graph, where each node divides the input dataset by some rules and provides new subdatasets. The terminal nodes called leafs contain the classified samples.

Table I
Geometric Features of Vector Shapes

Geometric Feature	**Equation**	**Description**
Geospatial area	$A = \sum p$	Each pixel *p* of LiDAR data projection is proportional to real geospatial area of Earth, therefore feature "area" is applicable for geospatial images.
Geospatial perimeter	$P = \sum b_p$	b_p is the external side length of border pixel. Many authors are against perimeter and perimeter-based features, because of coastline paradox. The modern most used resolution of DSM (digital surface model) is 1 m^2. If the constant resolution is accepted, this parameter is important for analysis.
Rectangularity [4], [6]	$R = \frac{A}{a \cdot b}$	*a* – major axis (length of minimal bounding rectangle) and *b* – minor axis (width of minimal bounding rectangle). The parameter describes the object shape similarity with rectangle shape.
Elongation [4]	$E_1 = 1 - \frac{b}{a}$	The character expresses how strong the shape is elongated.
Elongation [5]	$E_2 = \frac{2\sqrt{A}}{a \cdot \sqrt{\pi}}$	This feature is used to evaluate the elongation of basin shape, but it evaluates not only ratio of minor axis with major axis, it measure the circle/ellipse solidity.
Eccentricity [4]	$E_3 = \frac{\sqrt{a^2 - b^2}}{a}$	The ratio of distance between the ellipse focal and major axis.
Form factor [5]	$F = A / a^2$	Form factor is used in hydrology to analyse basin, it expresses the elongation of shape too.
Compactness [4-6]	$C_1 = \frac{P}{2\sqrt{\pi A}}$	The ratio between object area and circle area. Sometimes compactness is interchanged by the parameter circularity, which is equal to 1/C$_1$.
Compactness	$C_2 = P / A$	The ratio between perimeter and object area.
Convexity [4], [6]	$K = P_c / P$	P_c – perimeter of convex hull, where convex hull is the smallest ambient shape.
Solidity [4], [6]	$S = A / A_c$	A_c – area of convex hull. The solidity expresses the density of object - how many wholes the object contains.

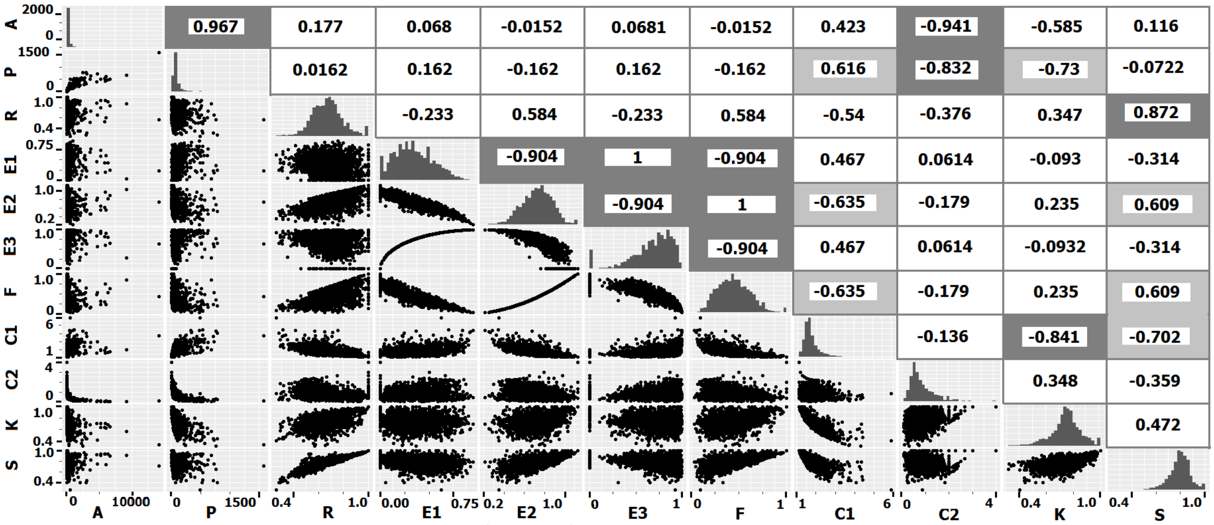


Fig. 3. Geometric features of building shapes and their correlation

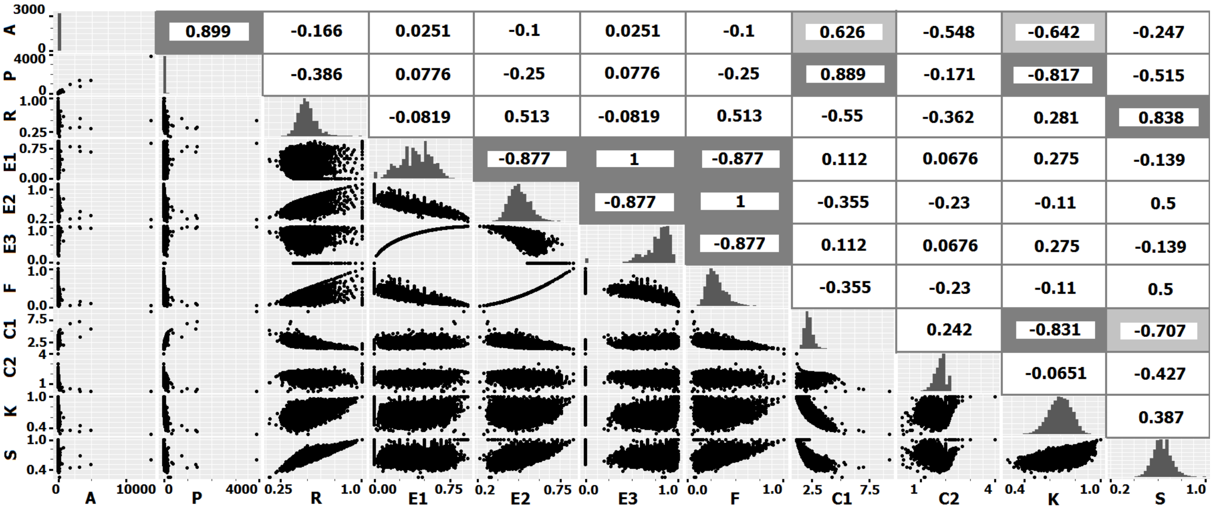


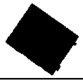
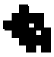
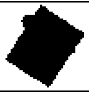
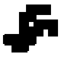
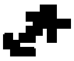
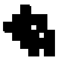
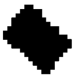
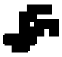
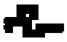
Fig. 4. Geometric features of noise shapes and their correlation

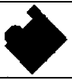
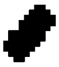
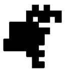
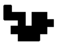
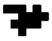
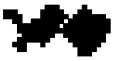
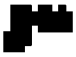
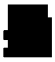
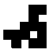
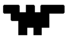
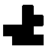
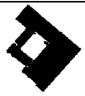
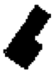
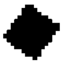
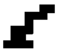
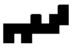
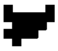
The common decision tree algorithms are *CART* (Classification and Regression Tree), *ID.3* (Interactive Dichotomizer 3) and *C4.5* [8].

Random forest is the ensemble method, which constructs many decision trees using bootstrap datasets with random feature set. The final prediction is defined by the majority of votes [7-8].

Random forests can be applied to measure the feature importance for object recognition. The most widely used score of importance of a given feature is the increasing in mean of the error of a tree (MSE for regression and misclassification rate for classification) in the forest, when observed values of this variable are randomly permuted in out-of-bag samples [7].

Table II
Geometric Features of Vector Shapes

	Building Shapes			Noise Shapes		
	1st quarter	Median	3rd quarter	1st quarter	Median	3rd quarter
A						
	26	70	139	13	16	21
P						
	28	48	72	24	28	36
R						
	0.59	0.67	0.75	0.43	0.48	0.54
E1						
	0.13	0.27	0.50	0.30	0.43	0.53
E2						
	0.64	0.75	0.85	0.53	0.59	0.66
E3						
	0.59	0.75	0.87	0.71	0.82	0.88
F						
	0.32	0.44	0.56	0.22	0.27	0.35
C1						

	1.48	1.63	1.85	1.87	2.04	2.28
C2						
	0.51	0.72	1.13	1.67	1.85	2.00
K						
	0.70	0.76	0.82	0.64	0.70	0.76
S						
	0.75	0.80	0.85	0.59	0.63	0.69

* If feature is marked by gray color, then buildings and noise objects don't have intersections of 1st and 3rd quarters or they are very small.

VI. GEOMETRIC FEATURE IMPORTANCE ANALYSIS

Feature importance is measured using the full dataset of unique shapes (buildings and noise objects). The selected number of trees is 500.

Completing the analysis of perimeter-based indices, the authors of scientific article [5] mention that some other authors suggest completely abandon shape indices because of fractal behave of boundary called coastline paradox. The most important feature is C_2 (see Fig.5), which is perimeter-based index and resolution dependant parameter (1 m^2 in this study). Therefore, it is a good argument, that resolution dependant indices must not be ignored, only they must be used and analysed considering the resolution of processing data.

Other features are important too, but they are not so strongly important as C_2 (see Fig.5), but diagram shows that the features {R, F} strongly changed their place after the correlated features had been removed, that coincides with the results of experiment mentioned in article [7].

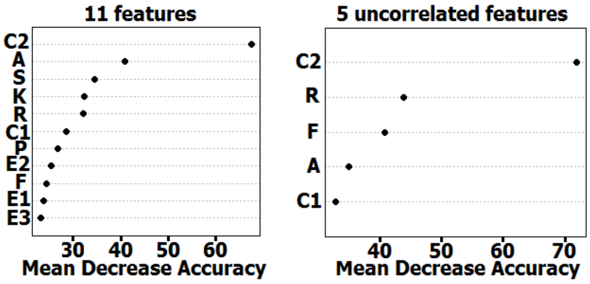


Fig. 5. Geometric feature importance: left diagram – 11 features, right diagram – selected 5 uncorrelated features

VII. VALIDATION OF RESULTS

Constructing the random forest for importance analysis 500 trees have been created, but 25 trees are sufficient number to classify object shapes (see Fig.6). Therefore 25 trees are used to complete the analysis of classification accuracy.

All dataset of unique shapes is randomly split into training set (80%) and validation set (20%). Four confusion matrices of classification are calculated (see Table III): two matrices for validation dataset (6997 shapes) using 11 features and only 5 uncorrelated features, dataset of unique shapes (34 992 shapes, where buildings and noise objects have 199 equal samples) and for raw dataset with all shapes (844 284 shapes).

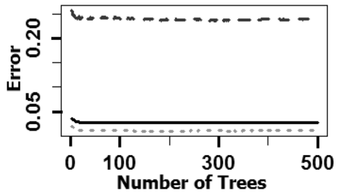


Fig.6. Dependence between error and number of trees in random forest

Table III
Confusion Matrices

	Validation dataset 11 features		Validation dataset 5 uncorrelated features	
	B	*N*	*B*	*N*
B	0.053	0.011	0.054	0.009
N	0.017	0.919	0.017	0.920
	A = 0.972	*K* = 0.776	*A* = 0.975	*K* = 0.800
	Unique Shapes		All Shapes	
B	0.062	0.004	0.003	0.784
N	0.009	0.925	0.000	0.213
	A = 0.987	*K* = 0.904	*A* = 0.216	*K* = 0.001

B – buildings, *N* – noise objects, *A* – total accuracy, *K* – Kappa coefficient

The confusion matrices show, that two models (11 features and 5 uncorrelated features classification models) do not have significant difference in accuracy.

The confusion matrix of unique shapes shows the strongly better accuracy, that can be explained, that the classification system remembered the shapes from training set, but it is more appropriate index in this case, because the shapes of buildings in most cases are similar.

The most interesting fact is real dataset recognition with very low accuracy and Kappa coefficient, which can be explained by "one pixel noise" mentioned in chapter III. "One pixel noise" is classified as building, because of square form of shape, which is very similar to building. Noise objects with small rectangle shapes provide the similar problem.

"One pixel noise" was selected by complex expression in chapter III, but noise filtering by area is more simple approach. The accuracy and Kappa coefficient increase after noise filtering by area, that is depicted in Fig.7.

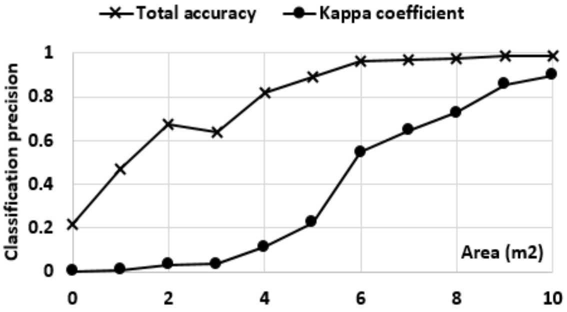


Fig.7. Classification accuracy depending on filtered objects by area

Completing filtering by 10 m^2 of area, 801 095 objects are removed, where 217 (0.03%) objects are buildings and 800 878 (99.97%) – noise objects. Remainder is classified by random tree, the accuracy is depicted in Table IV and sample in Fig.8.

Table IV
Confusion Matrix

A=0.99 *K*=0.90	Buildings	Noise
Buildings	0.050	0.005
Noise	0.006	0.939

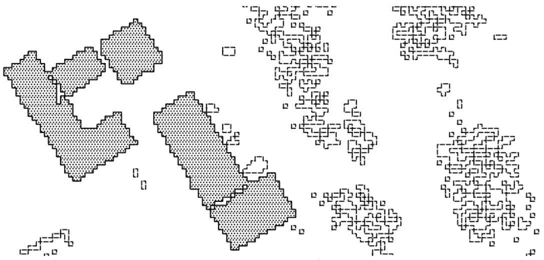


Fig.8. Buildings and filtered noise

VIII. RESULTS AND DISCUSSION

Eleven geometric features have been analysed under scope of this study. The correlation analysis reduced this number of features from eleven to five variables, where compactness index P/A (C_2) is the most important.

The distribution analysis showed that noise objects contains two groups of noises. One is similar to “salt and pepper” noise, which contains some pixel objects with square and rectangle shapes, for example, cars, poles and tree trunks. Other contains objects like bridges, walls or relatively big shapes of robust vegetation. The first group must be simply removed or ignored using area filter (area < 11 m^2), but the second group can be classified using geometric shapes (see Fig.7) with precision: total accuracy 0.99 and Kappa coefficient 0.90 (see Table III and IV).

IX. CONCLUSIONS

The geometric features have been analysed using the immediate output of 2D classification algorithm - the borders of shapes have toothed form. If line simplification algorithms are applied, the correlation and importance of features may be different.

The geometric feature “rectaliniarity” has not been analysed together with other features under scope of study, because it requires to use line simplification algorithms. Therefore it must be discovered independently to compare the best combination of algorithms and their input parameters with features researched under this study.

The combination of statistical, spatial and geometric features belong to different groups of parameters. Therefore correlation among them must be minimal, but clusters are located in sufficient distance one from other providing good conditions for automatic classification.

X. ACKNOWLEDGMENTS

Author expresses his gratitude to the State Land Service of Latvia and to Latvian Geospatial Information Agency for providing samples for the research purposes.

REFERENCES

[1] Y. Wang, L. Cheng, Y. Chen, Y. Wu and M. Li, “Building Point Detection from Vehicle-Borne LiDAR Data Based on Voxel Group and Horizontal Hollow Analysis,” *Remote Sensing*, vol. 8, no. 5, p. 25, May 2016 [Online]. Available: www.mdpi.com/2072-4292/8/5/419/pdf. [Accessed March 3, 2017].

[2] S. Kodors, A. Ratkevics, A. Rausis and J. Buls, “Building Recognition Using LiDAR and Energy Minimization Approach,” *Procedia Computer Science*. Vol. 3, pp. 109-117, December 2014 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187705091401583X. [Accessed March 3, 2017].

[3] C. Nesrine, G. Li and C. Mallet, “Airborne LiDAR Features Selection for Urban Classification Using Random Forests,” *Laser scanning 2009*, IAPRS, Vol. XXXVIII, pp. 207-212, September 2009 [Online]. Available: www.isprs.org/proceedings/XXXVIII/3-W8/papers/p207.pdf. [Accessed March 3, 2017].

[4] F.Y. Manik, Y. Herdiyeni and E.N. Herliyana, “Leaf Morphological Feature Extraction of Digital Image Anthocephalus Cadamba,” *TELKOMNIKA*, vol.14, no. 2, pp. 630-637, June 2016.

[5] A. Bardossy and F. Schmidt, “GIS approach to scale issues of perimeter-based shape indices for drainage basins,” *Hydrological Sciences-Journal-des Sciences Hydrologiques*, vol. 47, no. 6, pp. 931-942, December 2002 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.408.2607&rank=1. [Accessed March 3, 2017].

[6] N. Jamil and Z.A. Bakar, *Shape-Based Image Retrieval of Songket Motifs, Proceedings of the 19th Annual Conference of the National Advisory Committee on Computing Qualifications*, July 7-10, 2006, pp. 213-219.

[7] R. Genuer, J.-M. Poggi and C. Tuleau-Malot, “Variable selection using Random Forests,” *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225-2236, October 2010.

[8] S. Cinaroglu, “Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures,” *International Journal of Computer Applications*, vol. 138, no. 1, pp. 37-41, March 2016 [Online]. Available: www.ijcaonline.org/research/volume138/number1/cinaroglu-2016-ijca-908704.pdf. [Accessed March 3, 2017].