EVALUATING CLUSTERING METHODS ON TOPOGRAPHIC AND HIDROLOGICAL FEATURES ON LiDAR DATA AT FOREST ENVIRONMENT

Danilo Avancini Rodrigues[1]; Laurimar Gonçalves Vendrusculo[2]; Cornélio Alberto Zolin[3]; Tarcio Rocha Lopes[4].

[1] Graduando em Engenharia Florestal UFMT, Sinop, MT, danilo.avancini@gmail.com
[2] Dr., Pesquisador, Embrapa Informática Agropecuária, Campinas, SP, laurimar.vendrusculo@embrapa.br
[3] Dr., Pesquisador, Embrapa Agrossilvipastoril, Sinop, MT, cornelio.zolin@embrapa.br
[4] Engenheiro Agrícola, mestrando em Agronomia UFMT, Sinop, MT, tarcio281192@hotmail.com

INTRODUCTION

The acquisition of high resolution geographic data through laser technology has recently being expanded due to the development of LiDAR (*Light Detection and Ranging*) system. This technology's growth is relying on its great ability to acquire information in large quantity and short time. The geographic data provided from laser scanning is capable of raising information for coast planning, assess flooding risk, power transmission network and telecommunication, forests, agriculture, oil, transportation, urban planning, mining, among others (GIONGO et al., 2010).

LiDAR technology follows the same principles as the RADAR system, with the difference of using laser pulses to locate features, instead of radio waves. Not only for its ability to deal with large amounts of information in such a short period of time, LiDAR has the advantage upon the classic passive sensors (aerial photographs and satellite images) of not depending on a source of light, and so its data will never present shadows from clouds or neighboring features (GIONGO et al., 2010).

Data from LiDAR sensor is distributed in a point cloud where each point has at least three-dimensional spatial coordinates (latitude, longitude and height) that correspond to a particular point on the Earth's surface from which the laser pulse was reflected.

Once LiDAR data is acquired the next step is use algorithms that separate points (also referred to as returns) on the point cloud that represents the ground and the ones above the ground level, those algorithms can then process series of interpolation that allows the operator to generate Digital Elevation Models (DEMs). In order to add information for the points within the DEM, labeling those returns following a pattern and then grouping them on clusters is useful as one of the steps in exploratory data analysis.

Several methodologies were developed to organize a pattern of points in a multidimensional space into clusters based on similarity. Points belonging to the same cluster are given the same label and present a pattern where they are more similar to each other than they are to a pattern belonging to a different cluster (JAIN et al., 1999). One example to apply this technology on forestry activities is the application of silvicultural treatment to improve the forest's productivity, where the decision is taken considering characteristics from the site and sites with similar characteristics may have the same silvicultural system.

The variety of techniques for grouping data elements has produced a rich and often confusing assortment of clustering methods. Furthermore, there is a lack of studies grouping topologic and hydrologic variables at forested environments. The goal of this survey is to evaluate k-means and CLARA clustering techniques on a LiDAR-derived DEM from southern Amazonia, in the municipality of Cotriguaçu, Mato Grosso, Brazil.

MATERIAL AND METHODS

The LiDAR survey area used in this paper comprises 250 ha of secondary rain forest and lies in a region where weather prevails belonging to group A tropical rainy climate, type "Am", common to the short period of drought and rainfall below 60 mm in the driest period of the year. The vegetation covering the study area consist of Open Tropical Rain Forest, submontane formation with palm trees. Its physiognomy shows large trees spaced grouped with palm trees presenting a number of features with large leaves and rough bark (BRAZIL, 1982).

Esteio Ltda., a LiDAR data provider located in Curitiba, Paraná, Brazil, acquired the LiDAR data in one flight on 2011 using a Leica ALS-50 discrete-return LiDAR system. The LiDAR data used here is part of the data belonging to the Sustainable Landscapes Project, a project working through the cooperation of the Brazilian EMBRAPA and the United States Forest Service.

The LiDAR point cloud was summarized to create a digital elevation model (DEM). The DEM was created by first separating ground returns following the MCC (Multiscale Curvature Classification) algorithm from Evans and Hudak (2007). For the implementation of the MCC algorithm it is firstly required the definition of a vector Z which comprises the X coordinate, Y coordinate and Z (elevation) of all LiDAR returns. The Z vector is used to interpolate a raster surface using a thin-plate spline (TPS) interpolation at a cell resolution defined by scale parameter λ. A curvature tolerance (t) is then added to the interpolated surface and points are classified as nonground by applying the conditional statement "IF Z(s) > t THEN classify as nonground" (EVANS; HUDAK, 2007). In this study a scale parameter (λ) of 1.5 and a curvature tolerance (t) of 0.3 were applied to classify the returns as ground or nonground points.

This study estimated three topographic and hydrologic variables that are frequently associated to erosion and soil moisture condition. There variables are: slope, topographic wetness index (TWI). The slope tool of ESRI software ArcGIS 10.0 was applied to generate DEM map which calculate the maximum rate of change in degree between each cell and its neighbors, This approach provides a slope value for every cell in the output raster. Slope values will then represent a flatter terrain in case it is low, or steeper terrain if it is a high value.

In order to proceed for the Topographic Wetness Index (TWI) calculation (Equation 1), the flow accumulation needed to be calculated providing a flow accumulation value for every cell in the raster. Combining slope and flow accumulation, the TWI could then be calculated providing each cell in the output raster a TWI value that represent drainage depressions for higher values, and crests and ridges for lower values.

$$TWI = \ln (a \div \tan B) \quad (1)$$

where,
TWI – Topographic Wetness Index (TWI);
a – upstream contributing area (flow accumulation) (m^2) e
B – slope.

Through the R software environment for statistical computing (R CORE TEAM, 2015), k-means and CLARA (Clustering for Large Applications) (ROUSSEEUW et al., 2016) clustering algorithms were computed for every cell combining the data in a pattern of points

with similarity on Z (elevation), slope and TWI values. Finally a map with the four and six clustered was generated and evaluated using both clustering methodology.

RESULTS AND DISCUSSION

Exploratory analysis from means of variables set grouped by clusters and compared with same landscape position, as shown in Table I, reveled that means between both techniques were not significant different (Welch two sample t-test – $p < 0.05$). A clusters value from CLARA methodology was slightly higher than other clusters in K-means. However, considering only means attributes assessment it seems that both clustering brought any difference among clusters characteristics.

Table 1. Exploratory analysis.

Variable	K-means clustering				CLARA clustering			
	Cl 1	Cl 2	Cl 3	Cl 4	Cl 2	Cl 1	Cl 4	Cl 3
Mean Slope (degree)	17.4	5.4	10.8	5	15.6	5.5	10.8	5.1
Mean Elevation (m)	298	270	321	281	299	269	323	279
Mean TWI	2.9	3.9	3.3	3.9	3.04	3.9	3.3	3.9

The graphic representation of both clustering techniques k-means and CLARA with 4 and 6 clusters is illustrated on Figures 1 and 2. It was clear to notice that different labels were given to the clusters when individual methodology was applied. For example, the cluster representing crests, higher and steeper ground surfaces, which means low TWI and high Slope values were combined in the cluster labeled as 3 at the k-means clustering and as 4 at CLARA's when considering a total of 4 clusters (Figure 1). The same change can be observed when considering 6 clusters in total (Figure 2).

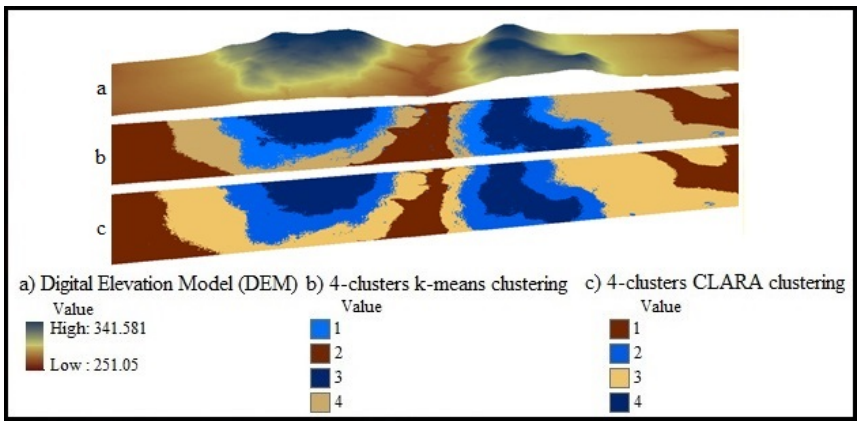

Figure 1. Graphical representation of the LiDAR-derived DEM (a), k-means (b) and CLARA (c) clustering with 4 clusters from the Cotriguaçu area.

The drainage depressions are clearer when using 6 clusters rather than 4 on both clustering methods. Among the two methodologies, visually, k-means shows itself more efficient when providing details about depressions on the ground surface.

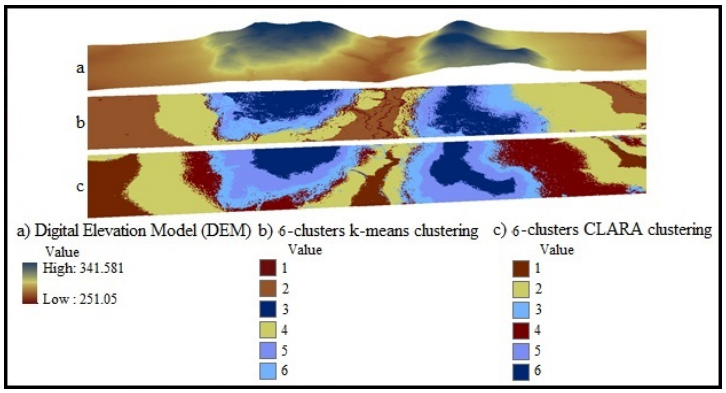

Figure 2. Graphical representation of the LiDAR-derived DEM (a), k-means (b) and CLARA (c) clustering with 6 clusters from the Cotriguaçu area.

When it comes to the area covered by each cluster, we can say that there is not a huge difference between k-means and CLARA for 4 clusters. But when considering 6 clusters, CLARA seems to overrate the area covered by clusters 2 to 5 over underestimating the area covered by clusters 1 and 6. In each method, the area corresponding to crests, or higher ground surfaces, are very close to each other in dimension. In the other hand, the area corresponding to the lower and flatter ground surface shows a great difference between the two methods.

CONCLUSION

The two clustering techniques provided different labeling for the same surface feature. K-means is a more efficient tool for providing details about drainage depressions. However, statistically speaking, it seems that clustering methodologies create clusters with similar characteristics. In both situations (4 and 6 clusters) the k-means method presented better imaging for topographic and hydrological features, being, thus, more recommended then CLARA when applying clustering to LiDAR-derived ground returns from forested areas.

REFERENCES

BRASIL. Ministério das Minas e Energia. **Levantamento dos recursos naturais**. [Rio de Janeiro: Ministério das Minas e Energia, 1982]. Project RADAMBRASIL. v. 20

EVANS, J. S.; HUDAK, A. T. A Multiscale Curvature Algorithm for classifying discrete return LiDAR in forested environments. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, n. 4, p. 1029-1039, 2007.

GIONGO, M.; KOEHLER, H. S.; MACHADO, S. A.; KIRCHNER, F. F.; MARCHETTI, M. LiDAR: princípios e aplicações florestais. **Pesquisa Florestal Brasileira**, v. 30, n. 63, p. 231-244, 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264-323, 1999.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2015.

ROUSSEEUW, P.; MAECHLER, M.; STRUYF, A.; HUBERT, M.; HORNIK, K.; STUDER, M.; ROUDIER, P. **Finding groups in data:** cluster analysis. [S.l.: s.n.], 2016. Available in: < https://cran.r-project.org/web/packages/cluster/cluster.pdf >. Access in: 10 may 2016.