# COMPARISON OF CLUSTERING TECHNIQUES APPLIED TO LASER DATA

H. Arefi[1]*, M. Hahn[1], F. Samadzadegan[2], J. Lindenberger[3]

[1]Dept. of Geomatics, Computer Science and Mathematics, Stuttgart University of Applied Sciences, Stuttgart, Germany
(hossein.arefi, michael.hahn)@hft-stuttgart.de
[2]Dept. of Geomatics, Faculty of Engineering, University of Tehran, Tehran, Iran, samadz@ut.ac.ir
[3]TopScan, Rheine, Germany, lindenberger@topscan.de

**Working Group IV/7**

**KEY WORDS:** LIDAR, First and Last Pulse data, Clustering, Quality assessment, Unsupervised learning, Feature Extraction

**ABSTRACT:**

During the last decade airborne laser scanning has become a mature technology which is now widely accepted for 3D data collection. Automated processes employ the scanned laser data and the platform orientation and other parameters of the scanning system to generate 3D point coordinates. These 3D points represent the terrain surface as well as objects on top of the terrain surface. Modern airborne LIDAR systems are able to record first pulse and last pulse range measurements together with the signal strength to provide more information about the reflecting surface or object.
The main goal of this paper is to investigate and compare procedures for clustering of LIDAR data. Classical clustering methods refer to a variety of methods that attempt to subdivide a data set into subsets or clusters. The study aims at a comparison of K-means clustering, competitive learning networks and fuzzy C-means clustering applied on range laser images. For comparison the confusion matrix concept is employed. The accuracy evaluation is done qualitatively and quantitatively. Experimental investigations are using LIDAR data taken from a scanning project in which the density of scanned points is around one point per square meter.

## 1. INTRODUCTION

Airborne laser scanning is an established technology for highly automated acquisition of digital surface models (DSM). Furthermore, in recent years LIDAR data has become as a highly acknowledged data source for interactive mapping of 3D man-made and natural objects from the physical earth's surface. The dense and accurate recording of surface points has encouraged research in processing and analysing the data to develop automated processes for feature extraction, object recognition and object reconstruction.

LIDAR data recoded with first and last pulse range and intensity values are considered the raw measurements in this paper. As these recordings are sets of irregularly distributed 3D points interpolation to a regular grid will ease processing of LIDAR data. It is well known that interpolation will low pass the raw data and thus some information will be lost. If the point density is high and almost regular this disadvantage will not be very significant. As our interest is in an investigation of clustering techniques using LIDAR data we assume the interpolation to be solved. The more interesting question in this regard is on the input for clustering algorithms. In addition to the LIDAR data, feature images reflecting texture and surface geometry are extracted and used for clustering. Clustering is a technique for image classification related to the unsupervised classification procedures. The overall goal is to extract information from the LIDAR data.

## 2. CLUSTERING TECHNIQUES

Clustering is a process of assigning pixels to categories or clusters based on some logic which acts on similarity of the pixels feature vectors.

Three clustering techniques which will be used in the experiments are described in the following. The clustering techniques are K-means (or hard C-means) clustering, fuzzy C-means clustering and competitive learning networks.

K-means is a representative for a classical and well explored unsupervised classification algorithm. Its counterpart in the fuzzy techniques is the fuzzy C-means algorithm which considers each cluster as a fuzzy set, while a membership function measures the possibility that each feature vector belongs to a cluster. Competitive learning networks pick up concepts of neural processing for unsupervised classification. The Competitive learning algorithm is based on a type of artificial neural network that possesses a self-organizing property called a simple competitive learning network.

### 2.1 K-means clustering algorithm:

K-means clustering, also known as hard C-means clustering, is one of the simplest unsupervised classification algorithms. The procedure follows a simple way to classify the data set through a certain number of clusters. The algorithm partitions a set of $n$ vector $X_j$ into $c$ classes $G_i$ , $i=1, \dots , c,$ and find a cluster centre for each class such that an objective function of dissimilarity, for example a distance measure is minimized. The objective

*Corresponding author

function that should be minimized, when the Euclidean distance is selected as dissimilarity measures can be described as:

$$J = \sum_{i=1}^{c} \left( \sum_{k, x_k \in G_i} \left\| x_k - c_i \right\|^2 \right), \quad (1)$$

where $\sum_{k, x_k \in G_i} \left\| x_k - c_i \right\|^2$ is the objective function within group *i*, and $\left\| x_k - c_i \right\|^2$ is a chosen distance measure between a data point $x_k$ and the cluster centre $c_i$.

The partitioned groups are typically defined by a $c \times n$ binary **membership matrix U**, where the element $u_{ij}$ is 1 if the *j*th data point $x_j$ belongs to group *i*, and 0 otherwise. That means:

$$u_{ij} = \begin{cases} 1 \text{ if } \left\| x_j - c_i \right\|^2 \le \left\| x_j - c_k \right\|^2, \text{for each } k \ne i, \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

The algorithm of K-means clustering is consists of the following steps:

*Step 1*: Initialize the cluster centres $c_i, i = 1, \ldots, c.$ This is typically achieved by randomly selecting *c* points from among all of the data.

*Step 2:* Determine the membership matrix U according to Equation (2).

*Step 3:* Compute the objective function according to Equation (1). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

*Step 4:* Update the cluster centres. Go to step 2.

The algorithm is significantly sensitive to the initial randomly selected cluster centres. The K-means algorithm can be employed multiple times to reduce this effect. More details can be found in (Jang 1997).

### 2.2 Fuzzy C-means clustering algorithm:

Fuzzy C-means clustering (FCM), also known as fuzzy ISODATA, is a data clustering algorithm in which each data point belongs to a cluster to a degree specified by its membership grade. Bezdek (1981, 1987) proposed this algorithm as an alternative to earlier (hard) K-means clustering.

FCM partitions a collection of n vector $x_i, i = 1, \ldots, \mathrm{n}$ into c fuzzy groups, and finds a cluster centre in each group such that an objective function of a dissimilarity measure is minimized. The major difference between FCM and K-means is that FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. In FCM the membership matrix U is allowed to have not only 0 and 1 but also the elements with any values between 0 and 1.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n. \quad (3)$$

The objective function for FCM is then a generalization of Equation (1) as following:

$$J(U, c_1, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{\,m} \left\| x_j - c_i \right\|^2, \quad (4)$$

Where $u_{ij}$ is between 0 and 1; $c_i$ is the cluster centre of fuzzy group *i*. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $u_{ij}$ and the cluster centres $c_i$ by:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{\,m} x_j}{\sum_{j=1}^{n} u_{ij}^{\,m}}, \quad (5)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\left\| x_j - c_i \right\|}{\left\| x_j - c_k \right\|} \right)^{2/(m-1)}}. \quad (6)$$

The FCM clustering algorithm is composed of the following steps:

*Step 1*: Initialized the membership matrix U with random values between 0 and 1 such that te constraints in Equation (3) are satisfied.

*Step 2*: Calculate c fuzzy cluster centres $c_i, i = 1, \ldots, c.$, using Equation (5).

*Step 3:* Compute the cost function (objective function) according to Equation (4). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

*Step 4:* Compute a new U using Equation (6). Go to step 2.

The cluster centres also can be initialized firstly and then iterative process carried out. (Jang 1997) provides a detailed behaviour of fuzzy c-means clustering including its variants and convergence properties.

### 2.3 Competitive Learning Networks

Competitive leaning networks are an unsupervised learning method which is based on the neural network concept. But unlike other neural learning algorithms in unsupervised classification no external teacher is available, thus only input vectors can be used for learning. In the following a competitive learning algorithm is described. Competitive learning is a self-organizing property of the neural network.

The method updates weights only on the basis of the input patterns. Figure (1) presents an example. All input units *i* re connected to all output units j with weight $w_{ij}$. The number of inputs is the input dimension, while the number of outputs is equal to the number of clusters. A cluster's centre position is specified by the weight vector connected to the corresponding output unit.

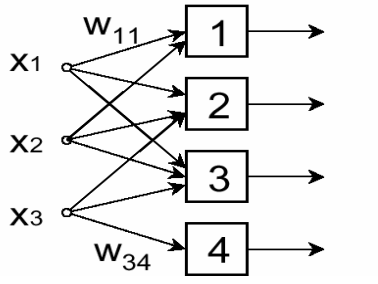


Figure 1: Competitive learning network

In Figure (1), the three dimensional input data are divided into four clusters and the cluster centres are updated via the competitive learning rule.

The input vector $x = [x_1, x_2, x_3]^T$ and the weight vector $w_j = [w_{1j}, w_{2j}, w_{3j}]^T$ for an output unit $j$ are generally assumed to be normalized to unit length. The activation value $y_i$ of output unit $j$ is then calculated by the inner product of the input and weight vectors:

$$y_j = \sum_{i=1}^{d} x_i w_{ij} = X^T W \qquad (7)$$

Where: $\|w_j\| = \sum_{i=1}^{d} w_{ij}^{\ 2} = 1$ for $j = 1, ..., M$ . Therefore we say that neuron $j^*$ is the winner of competition if: $y_{j^*} < y_j$ for all $j$, and $j^* = j$.

A simple competitive learning algorithm is composed of the following steps:

*Step 1*: Initialize all weights to random values and normalize them (so that $\|w_j\|$=1).

*Step 2:* Choose pattern vector $X$ from training set (input vector)

*Step 3:* Compute distance between pattern and weight vectors ($\|x_i - w\|$ ) and find the weights of the output with the smallest activation.

*Step 4:* Update the weight vector to:

$$w(t+1) = w(t) + \eta(t).(x_i - w(t)) \qquad (8)$$

*Step 5:* Go to step 2

Here $\eta(t)$ is monotonically decreasing in each iteration.

More details about the competitive learning method and its properties can be found in the Hung, Chih-Cheng (1993).

### 3. EXPERIMENTAL INVESTIGATIONS:

The airborne LIDAR data used in the experimental investigations have been recorded with TopScan's Airborne Laser Terrain Mapper system ALTM 1225 (TopScan, 2004). The data are recorded in a district called Ickern of the city of Castrop-Rauxel which is located in the west of Germany. The pixel size of the range images is one meter per pixel. This reflects the average density of the irregularly recorded 3D points which is fairly close to one per m2. Intensity images for the first and last pulse data have been also recorded and the intention was to use them too in the experimental investigations. Some first tests with these intensity images have been carried out but the current achievements have not yet been satisfactory. Figure (2) shows first- and last- pulse range images from the Ickern area. The impact of the trees in the first- and last- pulse images can be easily recognized by comparing the two images of this figure.

The first step in every clustering process is to extract the feature image bands. The features of theses feature bands should carry useful textural or surface related information to differentiate between regions related to the surface. Several features have

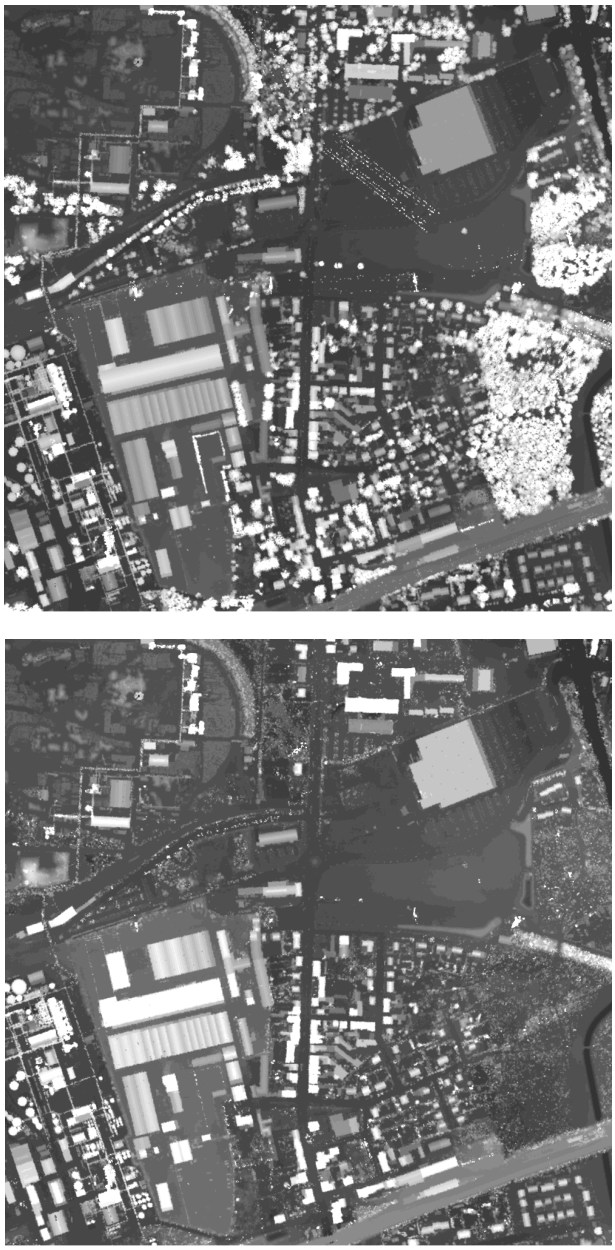

Figure 2: The first-pulse (above) and the last-pulse (below) range images.

been proposed for clustering of range data. Axelsson (1999) employs the second derivatives to find textural variations and Maas (1999) utilizes a feature vector including the original height data, the Laplace operator, maximum slope measures and others in order to classify the data. An investigation on LIDAR classification with remote sensing software packages was presented in Arefi et al., 2003).

In the following experiments we restrict to two types of features:

- The ratio between first and last pulse range images
- Top-Hat filtered last pulse range image

The normalized difference of the first and last pulse range images is used as the major feature band for discrimination of the vegetation pixels from the others. In analogy to the NDVI definition in Remote Sensing which is based on Red and NIR

channels of multispectral image data a range based NDDI is defined by

$$NDDI = \frac{fp - lp}{fp - lp} \quad (9)$$

where *fp* and *lp* indicate the first-pulse and last-pulse range image data, respectively.

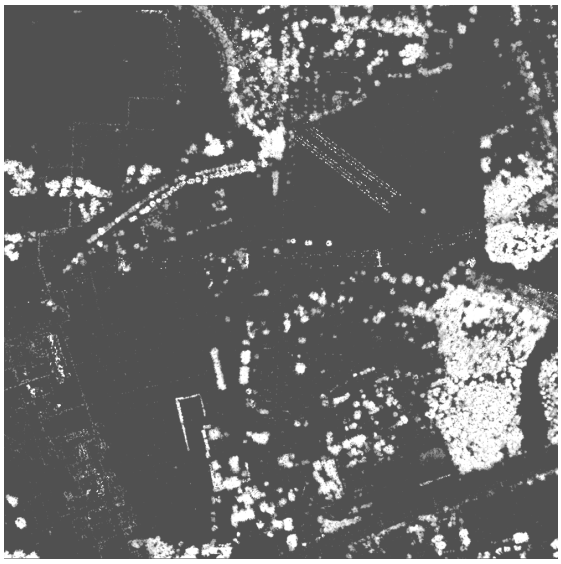

Figure 3: Normalized difference distance image derived from first-pulse and the last-pulse laser range images

Figure 3 shows that the *NDDI* image enhances vegetations areas with a significant 3D extend. In addition, it can be noticed in figure 3 that power lines show up in the *NDDI* image (Upper right region of Figure 3).

The morphology TopHat operator is utilized to filter elevation space. The TopHat transformation with a flat structuring element eliminates the trend surface of the terrain. A certain problem is to define the proper size of the structuring element which should be big enough to cover all 3D objects which can be found on the terrain surface. The TopHat operation is defined by:

$$TopHat = DSM - (DSM \circ se) \quad (10)$$

where DSM is the input surface for filtering, *se* is the structuring element function, and $\circ$ indicates the operator for grey scale opening morphology. The TopHat filtered last pulse range is shown in Figure 4. It enhances the 3D objects relative to the ground surface in the last pulse range image.

Input to the clustering processes is the *NDDI* ratio between first and last pulse range images (*NDDI* band) and the *TopHat* filtered last pulse range image (*TopHat* band). The three processes K-means clustering, fuzzy C-means clustering and competitive learning networks are employed and the results are shown in the following.

For all three clustering techniques we will

- restrict to four classes: a V-class, a B-class, a Background class and a Null class.

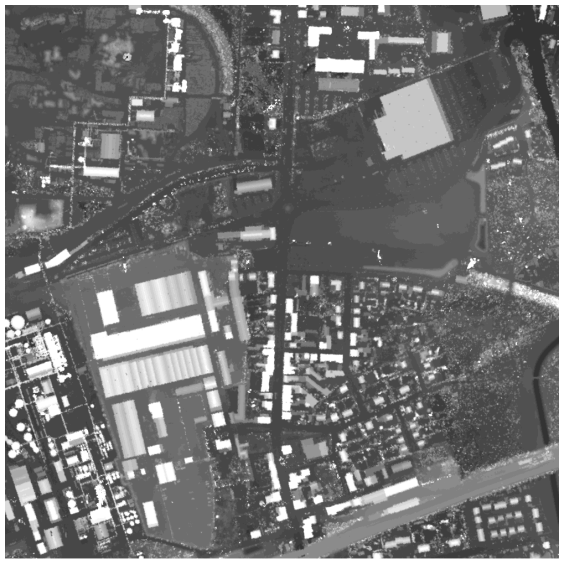

Figure 4: TopHat filtering of last-pulse laser range image

The term V – and B – classes are chosen because we expect that clustering based on the *NDDI* band and the *TopHat* band will directly point towards vegetation areas with significant 3D extend and building areas. But please note that this has no direct relation to supervised classification where training sets are selected and used for classification.

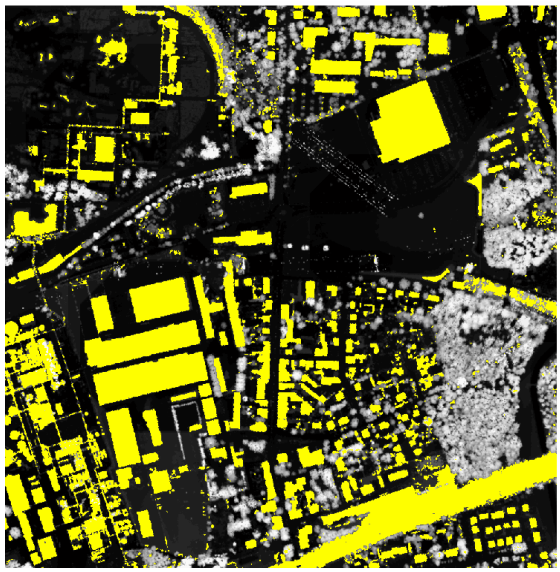

Figure 5: K-means clustering result (B class regions highlighted in yellow)

### 3.1 K-means clustering:

The K-means clustering results is shown in Figures (5) and (6). B class regions are highlighted in yellow in Figure 5 and V class regions in green colour in Figure 6. Visual inspections shows that V-class is directly associated with vegetation, in particular trees, bushes or forest and the B-class is mainly associated with building regions.

### 3.2 Fuzzy c-means clustering

Similarly utilizing the fuzzy C-means algorithm provides the results shown in Figures (7) and (8).

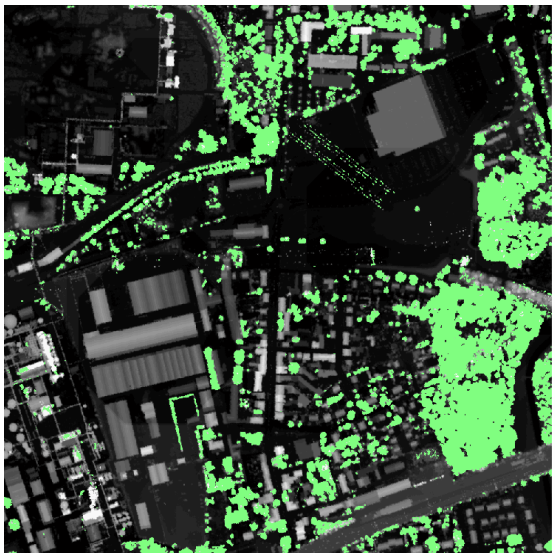

Figure 6: K-means clustering result (V class regions highlighted in green)

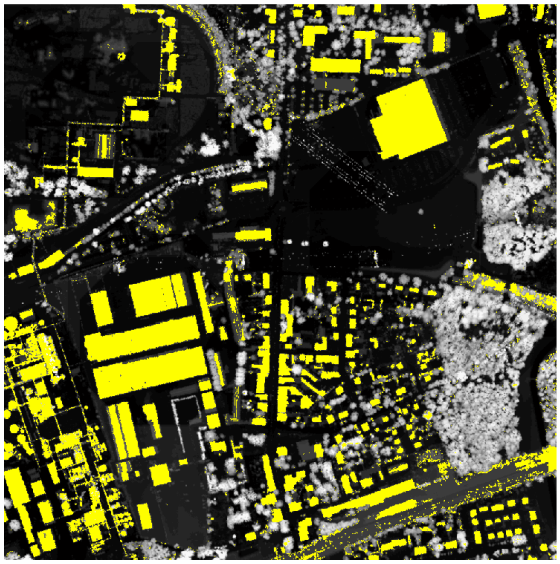

Figure 7: Fuzzy c-means clustering result (B class regions highlighted in yellow)

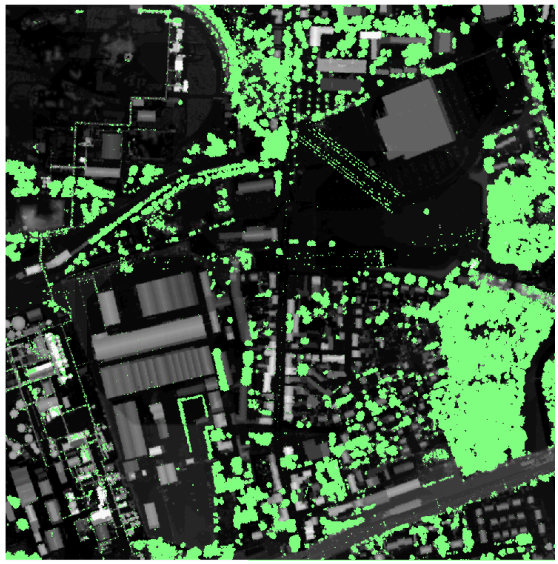

Figure 8: Fuzzy c-means clustering result (V class regions highlighted in green)

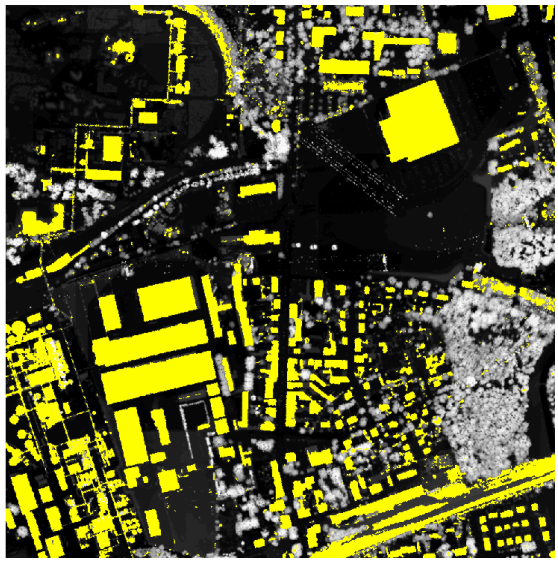

Figure 9: Competitive learning method (B class regions highlighted in yellow)

### 3.3 Competitive learning networks

The clustering results found by the competitive leaning networks algorithm are shown in Figures (9) and (10).

On a first view all three clustering algorithms provide reasonable classes which point back to vegetation, buildings and background. A comparison will be carried out in the next section. At this point we want to emphasis that with the two input channels *NDDI* band and *TopHat* filtered last pulse range image sufficient unique feature information is provided to clustering to separate the vegetation with 3D extend and building regions from background.

## 4. ANALYSIS OF CLUSTERING RESULTS

The confusion matrix is often used to discuss the results of image classification. Given some ground truth the relation between the "true" classes and the classification result can be quantified. With the clusters the same principle can be applied. Mostly a much bigger number of clusters is then related to the given ground truth classes to examine the quality of the clustering algorithm. If no ground truth is available the analysis may focus on comparing clustering results against each other. This kind of relative quality analysis is carried out in the following.

Input to the clustering processes has been the *NDDI* ratio between first and last pulse range images (*NDDI* band) and the *TopHat* filtered last pulse range image (*TopHat* band). The three processes K-means clustering, fuzzy C-means clustering and competitive learning networks are employed as discussed in the section before.

The following confusion matrix (Table 1) contains the number of pixels assigned to each cluster in the results of K-means clustering and competitive learning networks. Reading down a column shows how pixels in one class of K-means were assigned in the clustering results of competitive learning networks.

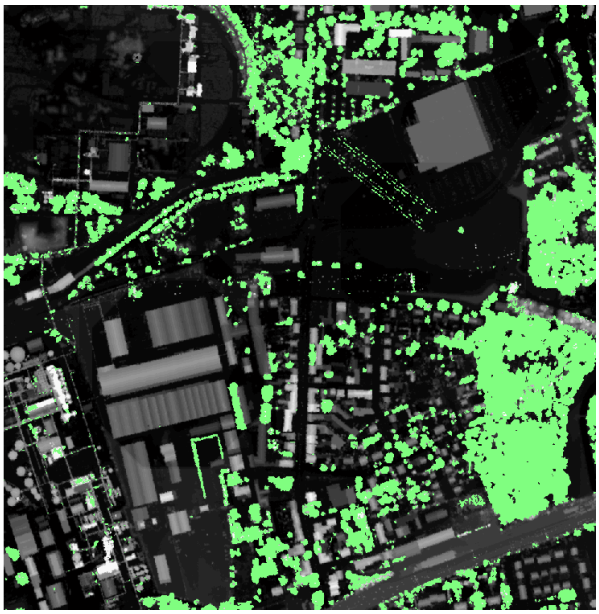

Figure 10: Competitive learning method (V class regions highlighted in green)

| | | K-means | | | |
|---|---|---|---|---|---|
| | | B-class | T-class | Back-ground | Null |
| Competitive learning networks | B-class | 178851 | 0 | 2140 | 0 |
| | T-class | 0 | 152557 | 7060 | 9945 |
| | Background | 25091 | 0 | 609439 | 0 |
| | Null | 769 | 45 | 6856 | 9248 |
| | Total | 204711 | 152602 | 625495 | 19193 |

Table 1: Confusion matrix between K-means clustering and Competitive learning method

Notice that the confusion matrix is almost diagonal which of course could be expected. It shows that both clustering algorithm recovered the three classes B-class, T-class and Background to a high degree of agreement. Null is used for rejection indicating assignment to none of the three classes.
In percentage values the degree of agreement between the clusters of both clustering algorithms is summarized in Table 2:

| Total common area | | |
|---|---|---|
| B-class | 180991 | 98,8 % |
| T-class | 169562 | 90,0 % |
| Background | 634530 | 96,1 % |
| Null | 16918 | 54,7 % |
| Total | 1002001 | 84,3 % |

Table 2: Common clustering areas of K-means clustering and Competitive learning networks method

That B-class and T-class can be easily identified with regions covered by buildings and trees was already discussed above.

| Class | K-means ( Count, % ) | Competitive learning ( Count, % ) | Fuzzy c-means ( Count, % ) |
|---|---|---|---|
| B-class | 180991, 18.1% | 204711, 20.4% | 144063, 14.4% |
| T-class | 169562, 17.0% | 152602, 15.2% | 196599, 19.6% |
| Background | 634530, 63.3 % | 625495, 62.4% | 657313, 65.6 % |
| Null | 16918, 1.7% | 19193, 1.91% | 4026, 0.40% |

Table 3: Clustering areas for all three clustering methods

Taking all three clustering areas simultaneously into account is shown in Table 3. Already by comparing the counts in each class a striking difference to the Fuzzy c-means result has to be observed. For the two classes of major interest in this study, the B-class and T-class, the differences are quite significant. Visual interpretation indicates that the B-class of K-means and competitive learning include building areas but also regions related to roads which supports the smaller number of counts of the fuzzy C-means method to be more precise. Similarly the higher number of counts for the T-class indication (3D) vegetation regions (trees, bushes) obtained with the fuzzy C-means method is supported by visual interpretation. Without ground truth we do not intend to draw further conclusion at this stage of our investigations.

## 5. SUMMARY

On a first view all three clustering algorithms provide reasonable classes which point back to vegetation, buildings and background. Comparison between the three clustering algorithms indicates a higher consistence of the results of K-means and Competitive learning networks. Fuzzy C-means deviates stronger but without comprehensive ground truth a absolute quality assessment is not feasible. The importance of the two input channels *NDDI* band and *TopHat* filtered last pulse range image for separating vegetation region with 3D extend and building regions from background has been shown clearly by the experiments.

## REFERENCES


Arefi, H., Hahn, M., Lindenberger, J., 2003. LIDAR data classification with remote sensing tools. Joint ISPRS Commission IV Workshop "Challenges in Geospatial Analysis, Integration and Visualization II", Stuttgart, September 8- 9.

Axelsson, P., 1999. Processing of laser scanner data – algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing,* 54(2-3): 138-147.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function. Plenum Press.

Bezdek, J.C., 1987. Some non-standard clustering algorithms. In: Legendre, P. & Legendre, L. *Developments in Numerical Ecology. NATO ASI Series, Vol. G14*. Springer-Verlag.

Hung, Chih-Cheng, 1993. Competitive Learning Networks for Unsupervised Training, International Journal of Remote Sensing, Vol. 14, No. 12, pp. 2411-2415.

Maas, H.G., 1999. The potential of height texture measures for the segmentation of airborne laserscanner data. *Proceedings of the Fourth International Airborne Remote Sensing Conference*, Ottawa, Canada. pp. 154-161.

TopScan, 2004. Airborne LIDAR Mapping Systems. http://www.topscan.de/en/luft/messsyst.html (accessed 10 Feb. 2004)