# AIRBORNE LASER SCANNING: CLUSTERING IN RAW DATA

Marco Roggero
Department of Georesource and Territory
Politecnico di Torino
Italy
roggero@atlantic.polito.it

Commission III, Working Group 3

**KEY WORDS**: Laser Scanning, clustering, DTM extraction, entity extraction.

## ABSTRACT

We implemented a strategy for terrain, vegetation and building detection, based on laser range data only. The result was obtained by working on raw data, so we were able to take advantage of the full resolution potential of laser scanning. The detection of objects was performed in two stages: first, elevated objects and ground are separated, and then the objects are classified as vegetation or buildings. Work is still in progress, about the extraction and classification of entities. A comparative analysis of the first pulse, the last pulse and intensity data can improve the result of clustering.
Results obtained in different environments with one-meter grid laser data are shown; we have tested the algorithm on city areas, countryside, river bed, landslides, mountains and wooded terrain.

## 1 INTRODUCTION

Laser scanning, provides high detailed Digital Surface Models, and we are able to extract a lot of information with very simple tecniques. A gridded network model is well suited for storage in a matrix and subsequently is well organized for simple computer algorithms. For example, a matrix can be represented as a raster, and we can use image processing tecniques to extract information. However, the nodes of the gridded network have to be constructed by interpolation in the original data set. Consequently some of the information will be lost. To take advantage of the full resolution potential of laser scanning, we must work on raw data, representing them, for example, by using irregular triangular or tetrahedral networks. Triangular models have been used in terrain modelling since the 1970s. However, because of limitations of computers and the complexity of TIN data structures, gridded models were preferred to triangular models.
We implemented a strategy to classify raw data using a simple data base structure, based on gridded networks. This strategy simplifies classification algorithms. Data base structure allows direct access to data; the algorithm assigns a flag and a cluster code to all the raw data, and this information is stored in the database. Then, the classified raw data can be represented as a triangular network. Detection of clusters in raw data is performed in two stages: first, elevated objects and ground are separated, and then the objects are classified as vegetation or buildings. The first stage is based on a local minimum criterion, which excludes elevated points from analysis. Then all the points within some distance of the estimated ground surface are classified as ground points. Ground evaluation is refined in two iterations. In the second stage the algorithm classifies non ground points as vegetation or buildings with a variance criterion.

## 2 TERRAIN EXTRACTION

The algorithm that extracts ground points from raw data, operates as a filter, applying a local operator to all the elements of the gridded network. The target of the first stage, is to obtain a rough approximation of the DTM, excluding elevated points. Algorithm calculates the minimum height in the local operator, then assigns points to the ground if their heights are compatible with a local slope; this idea is refined in the local regression criterion, that also considers the variance of the local data set. Then, the algorithm calculates height value in the center of the operator; this value is the weighted mean of the ground point heights in the operator. Weights depends on the distance from the operator center, normalized with Gauss distribution. Maximum building size affects local operator size.
In the second stage, the algorithm uses a threshold criterion to classify points as ground or non ground. Output of this classification is stored in the database. Finally the algorithm recalculates the DTM as the wheighted mean of ground points only.
These steps are repeated once. In the second iteration, to refine the classification, all the parameters (size of local operator, local slope, coefficients of variance propagation and threshold value) are more restrictive.

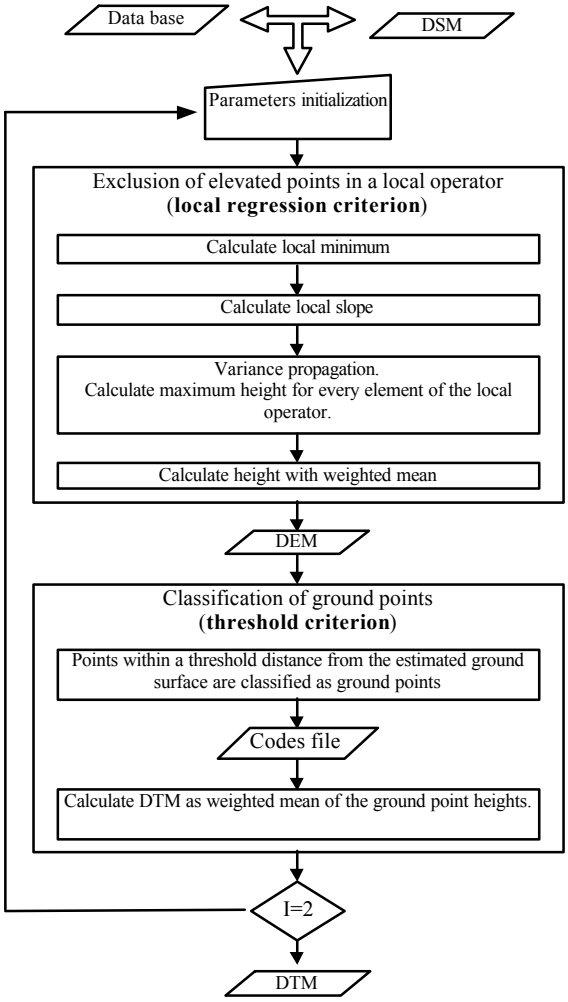


Figure 1: Algorithm flow chart.

**Classification criteria**

**Local regression criterion.** A criterion to detect ground points can be defined as a function of the maximum height difference between two points $p_i$ and $p_j$ in the local operator ($A$):

$$DTM = \left\{ p_i \in A \middle| \forall p_j \in A : h_{p_i} - h_{p_j} \le \Delta h_{\max}\left(d\left(p_i, p_j\right)\right) \right\} \quad [1.]$$

where $d$ is the planimetric distance [Vosselmann, 2000]. We have used this criterion, studing relations between maximum height difference and points distance. We can sort points by their planimetric distance from the local minimum $p_{min}$. To extimate a local slope, we calculate the parameters of the linear regression of the sorted data set

$$(x, y) = \left(d\left(p_i, p_{\min}\right), \left(h_i - h_{\min}\right)\right). \quad [2.]$$

This extimate assumes that points far from minimum affect the local slope less; so data are weighted with the factor $1/\sqrt[4]{d^2 + \Delta h^2}$ :

$$(x, y) = \left(d\left(p_i, p_{\min}\right)/\sqrt[4]{d^2 + \Delta h^2}, \left(h_i - h_{\min}\right)/\sqrt[4]{d^2 + \Delta h^2}\right) \quad [3.]$$

Linear regression of this population provides the parameters $a$ (intercept), $b$ (gradient), $s^2{}_a$ (standard deviation) and $s^2{}_b$ (gradient standard deviation). Parameters $a$ and $b$ and the propagation of their variance, define the local regression criterion, which links $Dh_{max}$ to the distance $d$:

$$\Delta h_{\max} = a + k_a^2 \cdot s_a^2 + b \cdot d + k_b^2 \cdot d^2 \cdot s_b^2 . \quad [4.]$$

The parameters $k^2{}_a$ and $k^2{}_b$ depends on terrain typology, so their calibration is very important. Indicative values are $k^2{}_a=5$ and $k^2{}_b=0.005$.

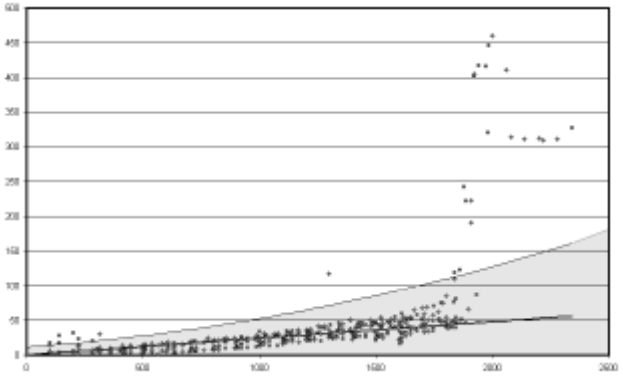

Figure 2: Linear weighted regression. Points in the grayed area are used to calculate the DTM.

**Threshold criterion.** The algorithm, after the computation of first approximation DTM on the basis of local regression criterion, classify raw data by vertical distance from the DTM. We define two threshold values, $k_1$ and $k_2$ (with $k_1<k_2$); the following contitions explains the classification criterion:

$h_i - h_{\min} < k_1/\cos b$ ground points

$h_i - h_{\min} > k_2/\cos b$ non ground points

$k_1/\cos b \le h_i - h_{\min} \le k_2/\cos b$ non classified points

Between $k_1$ and $k_2$, the algorithm is not able to classify points as ground or non ground. To restrict classification errors in zones with high slopes, threshold values are not constant, but depend on gradient.

## 3 CLASSIFICATION OF VEGETATION AND BUILDINGS

For many applications we need to detect the characteristics of scanned objects. This detection may be fully automatic. At first, classification of vegetation and buildings is very important. These two classes of objects are characterized by different values of variance in their spatial distribution, so we can use this parameter to extract vegetation and buildings from laser range data.

**Fitting a local plane.** Classifying the point $P(x_j,y_j,z_j)$, we consider it in relation to the N points $Q_i(x_i,y_i,z_i)$ with i=1,N, that satisfay the following condition:

$$d\left(P, Q_i\right) \le r \quad [5.]$$

where $r \in \Re$ is a parameter depending on sampling density, and $d\left(P, Q_i\right)$ is the euclidean distance

$$d(P, Q_i) = \sqrt{\left(x_P - x_{Q_i}\right)^2 + \left(y_P - y_{Q_i}\right)^2 + \left(z_P - z_{Q_i}\right)^2} . \quad [6.]$$

To fit the points $Q_i$ and the central point P with a local plane, their distance from the plane ax+by+cz =d has been minimized. The correction to the i-th point is:

$$v_i = z_i - \left(\frac{a}{c}x_i + \frac{b}{c}y_i - \frac{d}{c}\right). \quad [7.]$$

Its variance is

$$s^2_{v_i} = \frac{a^2}{c^2} s^2_{x_i} + \frac{b^2}{c^2} s^2_{y_i} + c^2 s^2_{z_i} \quad [8.]$$

The following function must be minimized:

$$c^2(a,b,c,d) = \sum_{i=1}^{N} \frac{v_i^2}{s_{n_i}^2} = \sum_{i=1}^{N} \frac{(ax_i + by_i + cz_i + d)^2}{as_{x_i}^2 + bs_{y_i}^2 + cs_{z_i}^2} \qquad [9.]$$

**Variance criterion.** In any 3D point distribution we can always fitting a local plane, if we have at least three points. But this plane may be or not be significant. We can accept or reject the hypothesis that the local plane is significant, by $\chi^2$ test on the four parameters a, b, c and d, or on any single parameter.
Then, Student's t test applied to normalized residuals tell us if the central point fits the local plane.
So, we can observe the following conditions:

1 The local plane exist, and the central point P fit the plane.
2 The local plane exist, but the central point don't fit the plane.
3 The local plane don't exist.

Classifying objects with this criterion, we can assume that in the case 1 the point belong to a building, or to another artificial structure, in the case 2 the point is an outlier, and in the case 3 belong to the vegetation.

**Raw data classification.** We have performed two different classifications in raw data, using local regression and threshold criteria, and using variance criterion. Now we can join the output of these two classifications, associating a decision to every couple of results. For example, if a point is classified as non ground by local regression and threshold criteria, and moreover fit a local plane, we can decide to assign it to a building. The decisions are described in table 1.

| Output of local regression and threshold criteria | Output of variance criterion | Combination of criteria |
|---|---|---|
| Non ground | Fit the local plane | Building |
| | Don't fit the local plane | Outlier |
| | Don't exist a local plane | Vegetation |
| | Isolated point | Outlier |
| Not classified | Fit the local plane | Ground |
| | Don't fit the local plane | Outlier |
| | Don't exist a local plane | Vegetation |
| | Isolated point | Outlier |
| Ground | Fit the local plane | Ground |
| | Don't fit the local plane | Outlier |
| | Don't exist a local plane | Rough ground |
| | Isolated point | Ground |

Table 1 – Decisions associated to every combination of classification results.

## 4 TESTS

We obtained DTMs in zones with very different morphology. The algorithm was tested in built-up areas, countryside, river beds, landslides, mountains and wooded areas. We performed tests on four data sets, acquired with three different scanning systems and in different sampling modes. So, the algorithm was tested in a large range of situations and applications.

**TopoSys data set on Pavia town.** Historical city of Pavia is a hard test area for every classification algorithm. Narrow streets and complex buildings alternate with parks and gardens, and vegetation is often close to buildings. Other sources of noise are car parks!
Table 4 report tests results. Tests 1 and 2 concerning flight T1 were performed on the same zone of Pavia city with different parameters. In this area there are no steep slopes, and the more restrictive parameters used in test 2 gave best results. Area used in test 3 has steep slopes near the castle; so we used less restrictive parameters that those used in test 1.

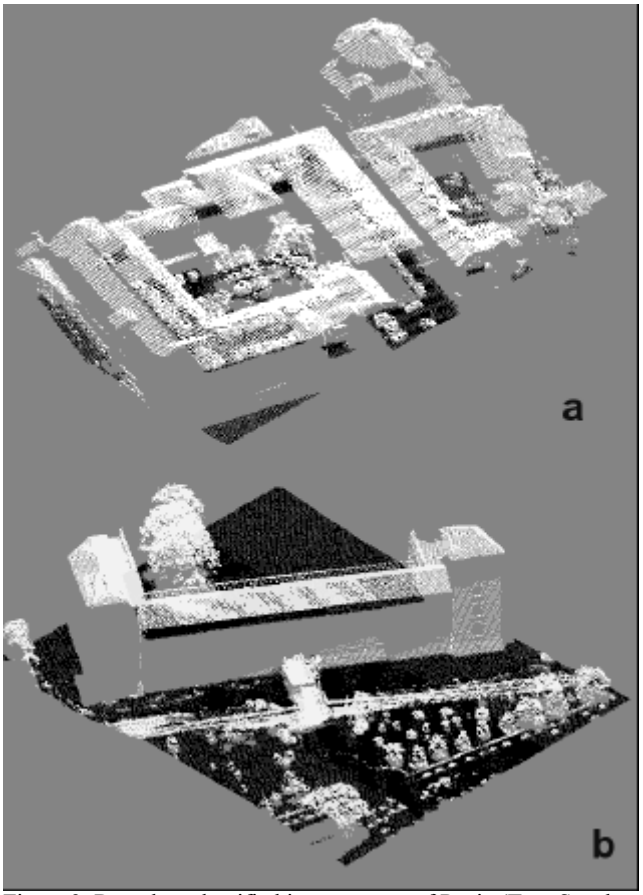


Figure 3: Raw data classified in two zones of Pavia (TopoSys data set). a) Pavia city, ground is correctly classified; here we have performed test 1 and 2. b) Castle, there are some errors in ground classification near castle moat; this is the area of test 3.

**TopoSys data set on Corniglio landslide.** Test zone is mountainous and densely wooded. DTM extraction is very difficult, also because we have first pulse data only. Wooded areas are very vaste, so operator size must be greater than in other cases. In tests we used a size of 41 m, but this size was not sufficient.

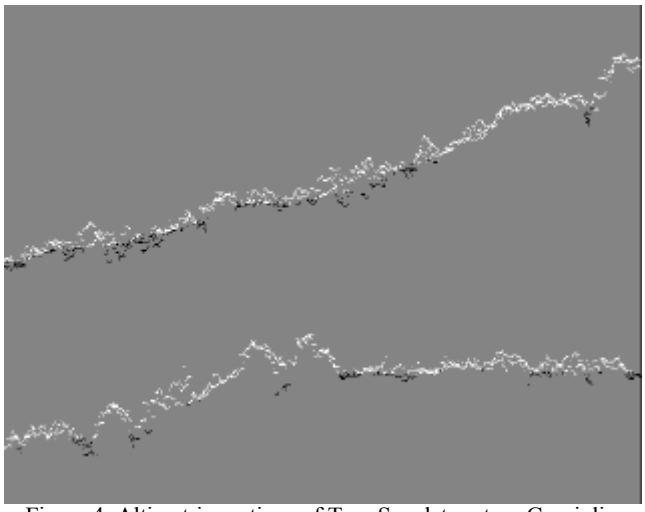

Figure 4: Altimetric sections of TopoSys data set on Corniglio landslide. Points classified as ground are in black, as non ground in white.

**TopEye data set.** This data set is very simple to process. Scanning system has sampled last pulse and first puls together, on a poorly wooded area. Besides, last pulse measurements has a good penetration of vegetal canopy, and give a regular representation of ground. These good conditions permit to calculate DTM classifying, in every knot of regular network, height measurements respect to median $\breve{m}$ . We define DTM as:

$$DTM = \{\overline{p}_i \in A : (\breve{m} - R) \le p_i \le (\breve{m} - r)\} \quad [10.]$$

if $(\breve{m} - R, \breve{m} - r)$ is not empty

$$DTM = \{\overline{p}_i \in A : (\breve{m} - r) \le p_i \le (\breve{m} + r)\} \quad [11.]$$

if $(\breve{m} - R, \breve{m} - r)$ is empty

where r and R are two threshold values.

We have calculated DTM with the iterative algorithm too. The data set describes the ground without gaps, so the algorithm doesn't need very large operator size. Variance propagation parameters instead, need greater than average values; this is due to steep slopes ground in many zones.

The difference between the two DTMs is a few centimeters, but it is greater on slopy ground.

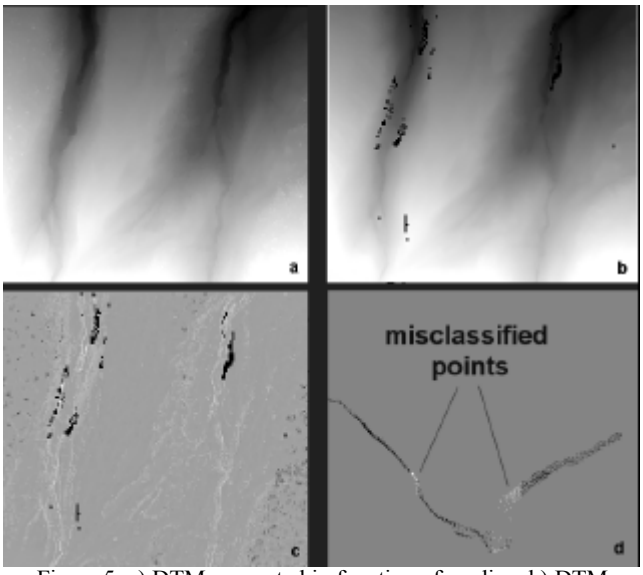


Figure 5: a) DTM computed in function of median. b) DTM calculated with iterative refinements. The zones in which the algorithm is not able to classify ground points are in black; these zones are vertical o sub-vertical ground. c) Difference between the two DTMs has an average of 7 cm and an RMS of 20 cm. d) Misclassified points.

The algorithm is not able to classify ground points correctly, if slope is vertical or sub-vertical, as you can see in figure 5d. This error has been limited by using threshold value function of slope (see threshold criterion). In the Bracigliano test area, misclassified points are 1% of total; this value may increase in mountainous areas with very uneven terrain.

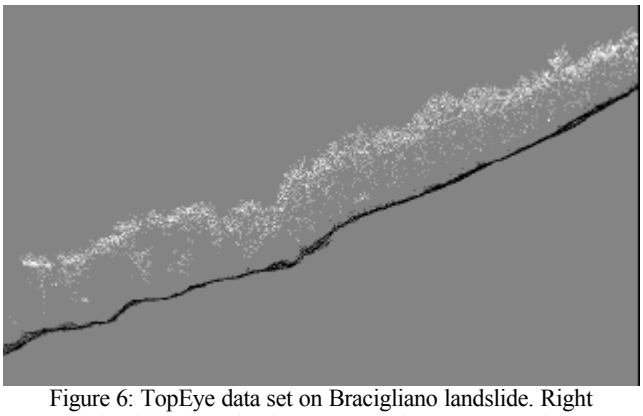

Figure 6: TopEye data set on Bracigliano landslide. Right classification of points as ground or non ground.

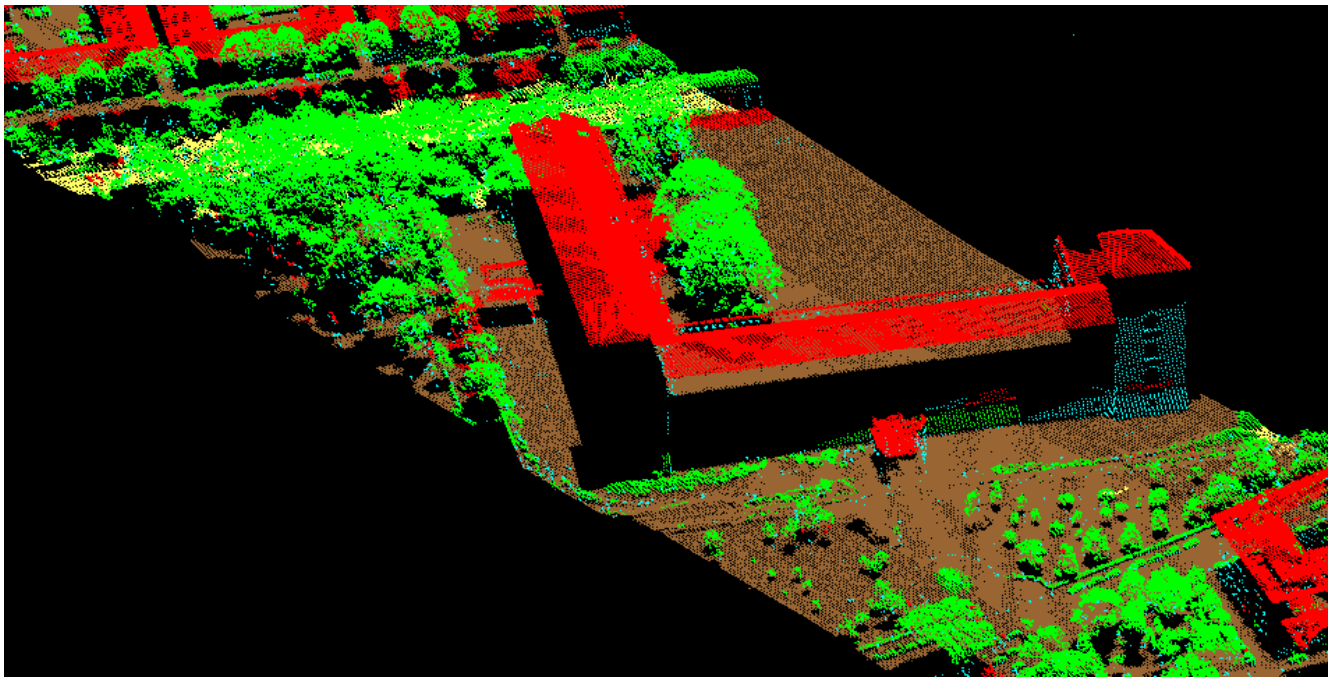

Figure 7 – Test on Pavia town, the castle (• ground, • rough ground, • vegetation, • buildings, • outliers).

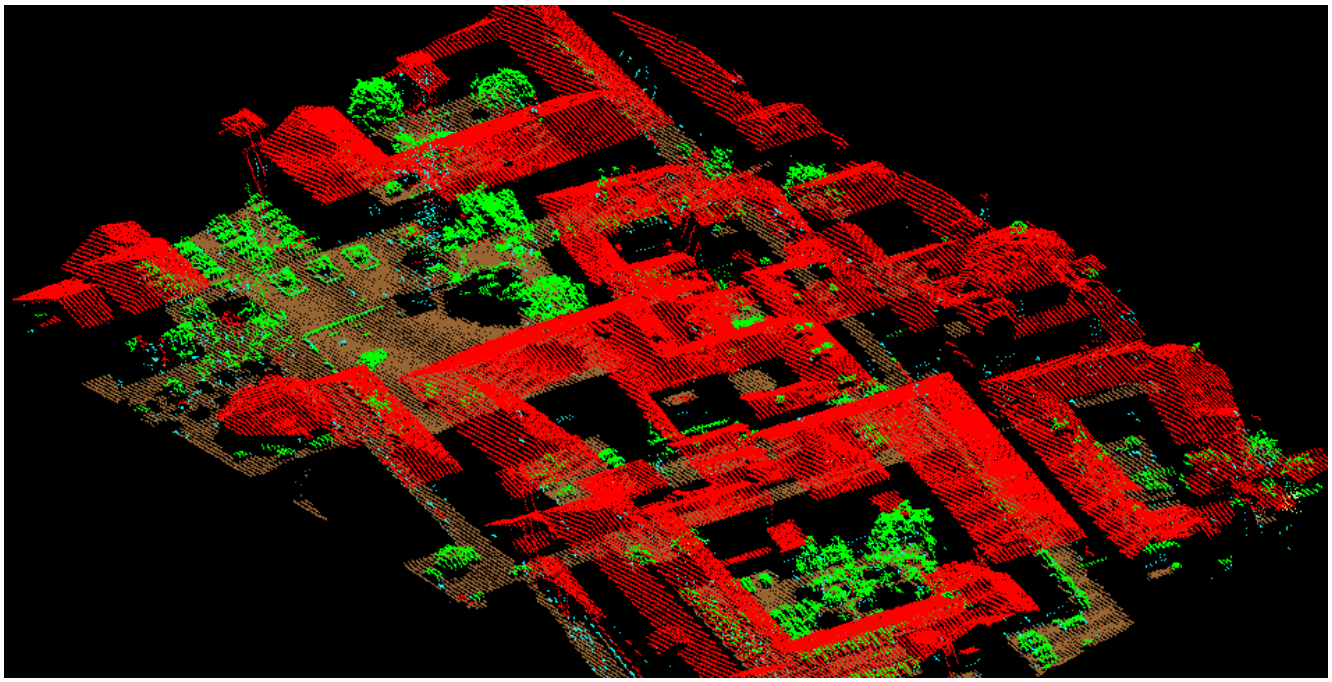

Figure 8 - Test on Pavia town, the medieval towers and the old city (• ground, • rough ground, • vegetation, • buildings, • outliers).

| System | Fly zone | Fly | Frequence [kHz] | Height [m] | Sampling density [punti/m$^2$] | Acquisition |
|---|---|---|---|---|---|---|
| TopoSys | Pavia town: historic city, countryside, industrial area, Ticino river. | T1 | 80 | 400 | 11,6 | L.P. |
| | | T2 | 80 | 850 | 5,5 | L.P. |
| | | T3 | 80 | 850 | 5,5 | F.P. |
| TopoSys | Corniglio landslide: mountain area; densly wooded area. | T4 | 80 | 300 ÷ 900 | 5 ÷ 13 | L.P. |
| Optech | Pavia town: historic city, countryside, industrial area, Ticino river. | O1 | 10 | 600 | 1,0 | F.P. & L.P. |
| | | O2 | 10 | 650 | 0,4 | F.P. & L.P. |
| | | O3 | 10 | 500 | 2,8 | F.P. & L.P. |
| TopEye | Bracigliano landslide: mountain area; wooded area. | E1 | | 100 ÷ 200 | 17 | F.P. & L.P. |

Table 2: Data sets used in tests.

| System | Fly | Test | First iteration parameters | | | | Second iteration parameters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Operator size | Regression | | Threshold | Operator size | Regression | | Threshold | |
| | | | O | ka | kb | k1 | O | ka | kb | k1 | k2 |
| TopoSys | T1 | 1 | 41·41 | 10 | 0,01 | 100 | 11·11 | 5 | 0,005 | 50 | 100 |
| | T1 | 2 | 41·41 | 5 | 0,005 | 50 | 11·11 | 2 | 0,002 | 25 | 50 |
| | T1 | 3 | 41·41 | 5 | 0,005 | 50 | 11·11 | 2 | 0,002 | 25 | 50 |
| TopoSys | T4 | 1 | 41·41 | 20 | 0,02 | 200 | 11·11 | 10 | 0,01 | 100 | 200 |
| | | 2 | 41·41 | 10 | 0,01 | 100 | 11·11 | 5 | 0,005 | 50 | 100 |
| TopEye | E1 | 1 | 11·11 | 20 | 0,02 | 200 | 5·5 | 10 | 0,01 | 100 | 200 |

Table 3: Parameters values assumed in tests.

| System | Fly | Test | Classified points [%] | | | Misclassified points [%] | | | Not classified points |
|---|---|---|---|---|---|---|---|---|---|
| | | | Ground | Vegetation | Buildings | Ground | Vegetation | Buildings | |
| TopoSys | T1 | 1 | 29,9 | 67,7 | | 0,0 | 1,6 | | 2,4 |
| | | 2 | 27,0 | 71,9 | | 0,0 | 0,0 | | 1,1 |
| | | 3 | 48,0 | 49,2 | | 0,0 | 5,4 | | 2,8 |
| TopoSys | T4 | 1 | 41,4 | 43,9 | - | ~30 | 0,0 | - | 14,8 |
| | | 2 | 17,8 | 75,8 | - | ~7 | 0,0 | - | 6,3 |
| TopEye | E1 | 1 | 88,6 | 10,4 | - | 0,0 | 1,0 | - | 0,9 |
| Values in this table are indicative, because referred to test areas, not to total fly. | | | | | | | | | |

Table 4: Sintetic results of tests.

## 5 CONCLUSIONS

The most important result of the tests is that the algorithm we have proposed is very sensitive to the parameters. This fact has a negative influx on algorithm output. But algorithm has classified points successfully, if parameters were well calibrated. Safely algorithm require parameter calibration on test areas, before proceeding to whole data set processing. A comparison with other algorithms will be very useful.

We used results obtained in this study as auxiliary knowledge for data segmentation.

Algorithm is implemented in DSM_Laser software.

## REFERENCES

[Agrawal, 1998] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, 1998. Automatic subspace clustering of high dimensional data for data mining applications, San Jose, CA.

[Barbarella, 2001] M. Barbarella, C. Fazio, 2001. Surveying of zones at risk of landslide by laser scanning. In: OEEPE Workshop on Airborne Laser Scanning and Interferometric SAR, Stockholm.

[Forlani, 2001] G. Forlani, C. Nardinocchi, 2001. Detection and segmentation of building roofs from LIDAR data. In: Workshop on 3D Digital Imaging and Modeling, Padova.

[Roggero, 2001] M. Roggero, 2001. Dense DTM from laser scanner data. In: OEEPE Workshop on Airborne Laser Scanning and Interferometric SAR, Stockholm.

[Vosselmann, 2000] G. Vosselmann, 2000. Slope based filtering of laser altimetry data. In: International Archives of Photogrammetry and Remote Sensing, Vol. XXXIII, Amsterdam.