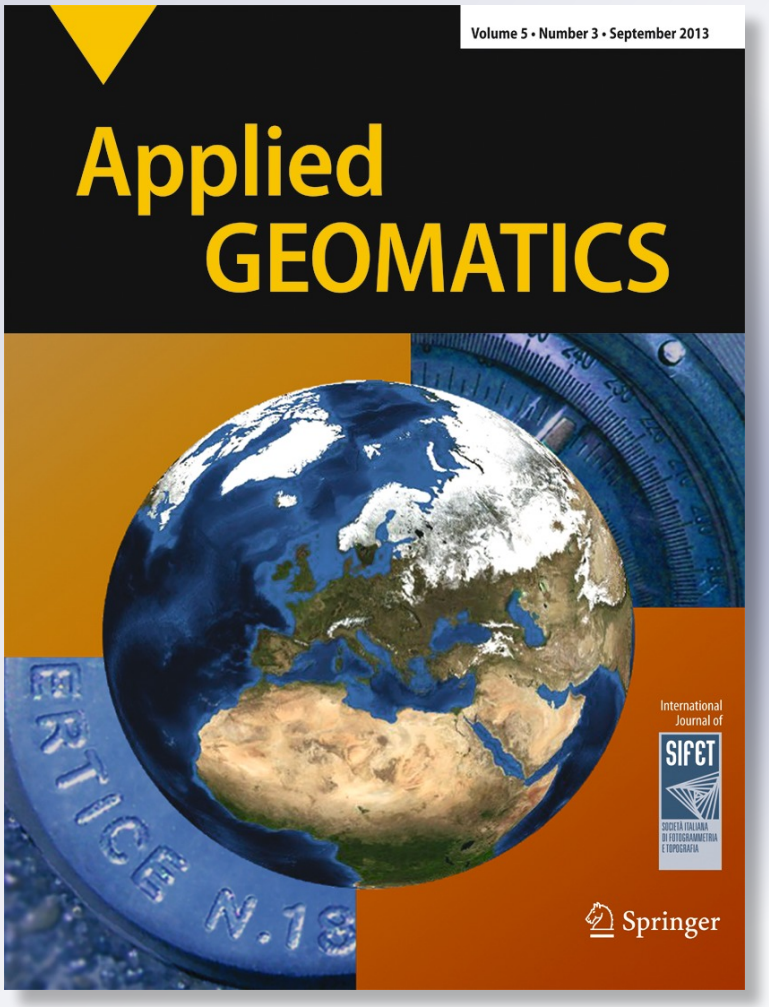
# *LiDAR data filtering and classification by skewness and kurtosis iterative analysis of multiple point cloud data categories*

**Fabio Crosilla, Dimitri Macorig, Marco Scaioni, Ivano Sebastianutti & Domenico Visintini**

**Applied Geomatics**

ISSN 1866-9298
Volume 5
Number 3

Appl Geomat (2013) 5:225-240
DOI 10.1007/s12518-013-0113-9




Springer

Appl Geomat (2013) 5:225–240
DOI 10.1007/s12518-013-0113-9

ORIGINAL PAPER

# LiDAR data filtering and classification by skewness and kurtosis iterative analysis of multiple point cloud data categories

**Fabio Crosilla · Dimitri Macorig · Marco Scaioni · Ivano Sebastianutti · Domenico Visintini**

Received: 12 January 2012 / Accepted: 2 April 2013 / Published online: 20 April 2013


**Abstract** A new procedure supporting filtering and classification of LiDAR data based on both elevation and intensity analysis is introduced and validated. After a preliminary analysis to avoid the trivial classification of homogeneous datasets, a non-parametric estimation of the probability density function is computed for both elevation and intensity data values. Some statistical tests are used for selecting the category of data (elevation or intensity) that better satisfies a bi- or a multi-modal distribution. The iterative analysis of skewness and kurtosis is then applied to this category to obtain a first classification. At each step, the point with the highest value of elevation (or intensity) is removed. The classification is then refined by studying both statistical moments of the complementary data category, in order to look for potential sub-clusters. Remaining clusters are identified by applying the same iterative procedure to the still unclassified LiDAR points. For more complex point distribution shapes or for the classification of large scenes, a progressive analysis is proposed, which is based on the partitioning of the entire dataset into more sub-sets. Each of them is then independently classified by using the core procedure. Some numerical experiments on real LiDAR data confirmed the potentiality of the filtering/classification method.

**Keywords** LiDAR · Filtering · Classification · Algorithms · Skewness · Kurtosis

This research has been partially presented at the ISPRS Workshop Laser Scanning 2011, 29–31 August 2011, Calgary (Canada), and published in Crosilla et al. (2011).

F. Crosilla (✉) · D. Macorig · D. Visintini
Department of Civil Engineering and Architecture, University of Udine, via delle Scienze 206,
33100 Udine, Italy
e-mail: fabio.crosilla@uniud.it

D. Macorig
e-mail: makkerig_dd@hotmail.com

D. Visintini
e-mail: domenico.visintini@uniud.it

M. Scaioni
Centre for Spatial Information Science and Sustainable Development Applications, College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road,
200092 Shanghai, People's Republic of China
e-mail: marco@tongji.edu.cn

I. Sebastianutti
Municipality of Tavagnacco (UD), Piazza Indipendenza 1,
33010 Feletto Umberto, Udine, Italy
e-mail: i.sebastianutti@comune.tavagnacco.ud.it

## Introduction

*Filtering* and *classification* of airborne LiDAR (light detection and ranging) data is a complex task, requiring the exploitation of different properties. Several approaches have been developed up until now, although no one seems to feature a general applicability.

*Filtering* methods refer to discarding all off-terrain points in order to reconstruct the bare-earth only. As proposed in Vosselman and Mass (2010), filtering methods can be grouped in four classes: *morphological filtering* (Haralick and Shapiro 1992), *progressive densification* (Axelsson 2000), *surface-based filtering* (Kraus and Pfeifer 1998), and *segment-based filtering* (Sithole and Vosselman 2005). In Sithole and Vosselman (2004), a comparison among different techniques applied to the same benchmarking dataset can be found.

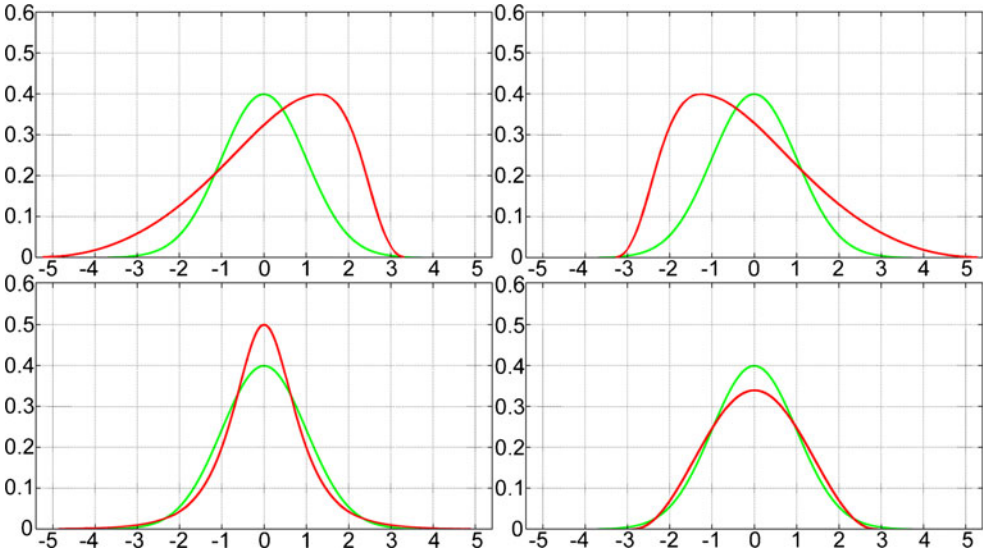


**Fig. 1** On the *top*, two examples of PDFs (in *red lines*) with skewness smaller (*left*) and larger (*right*) than a normal PDF (in *green line*). On the bottom, two examples of PDFs (in *red lines*) with kurtosis larger (*left*) and smaller (*right*) than 3, which corresponds to normal PDF (in *green line*)

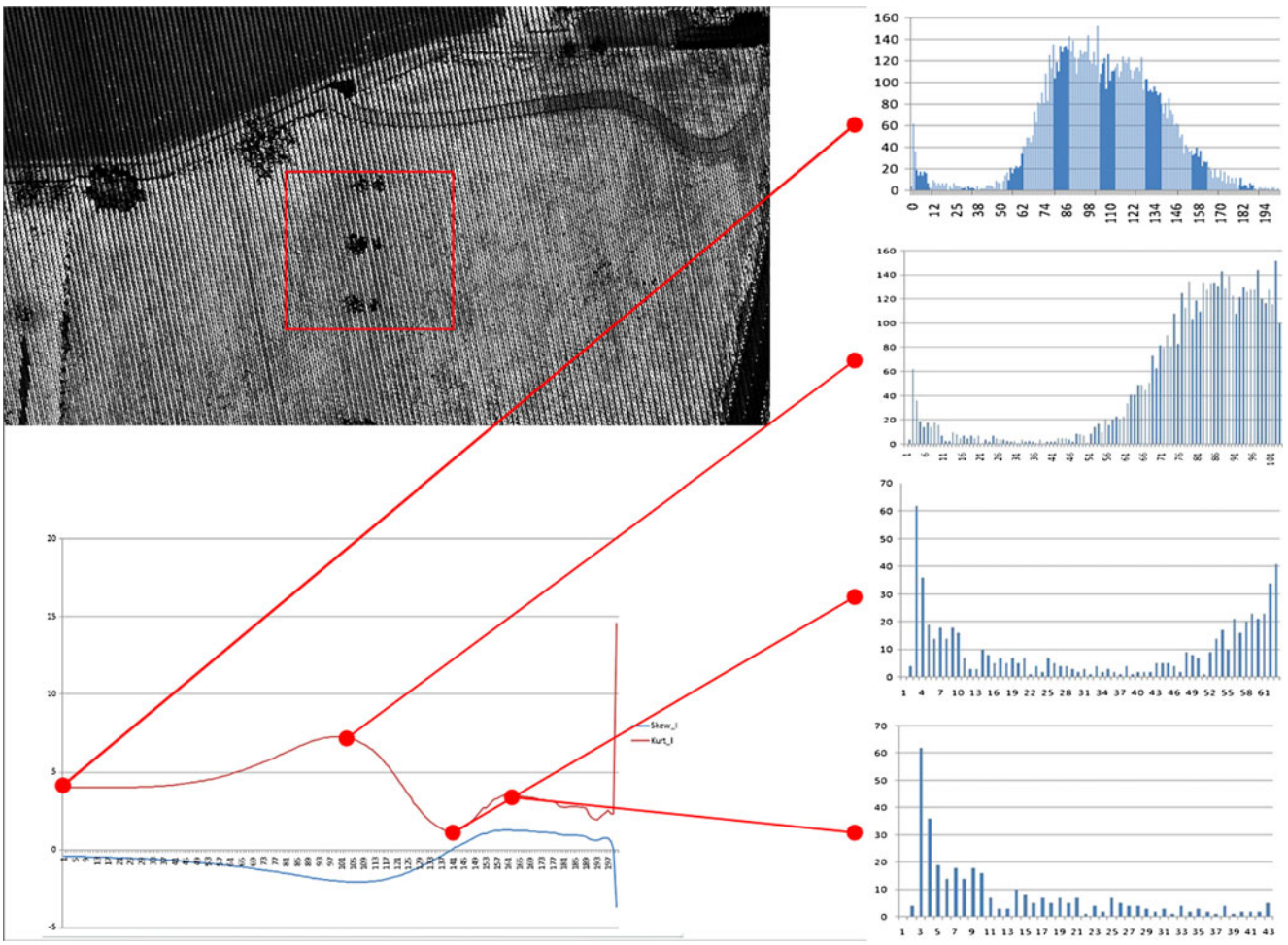


**Fig. 2** Example of classification of a *bi-modal* dataset of airborne LiDAR data. The analyzed area is shown with a *red contour* in the *intensity image* (*upper-left sub-figure*). The *right column* reports the histograms corresponding to four different steps of iterations along with the point featuring the highest intensity value (*lighter points* in the *intensity image*, corresponding to ground) is removed at each step. At the *bottom-left*, plots of skewness (in blue line) and kurtosis (in red line) computed along iterations are shown

*Classification* methods are used to split different categories of objects into homogeneous classes. Different classes are considered here, not only the bare terrain as in filtering procedures. Among these methods, some exploit the integration of multiple properties, chiefly *elevation* and laser *intensity*. This solution could increase the chance to accomplish a successful classification in those regions where a single data type does not provide enough information content. In the same direction, sometimes also aerial images are used to support the classification process (Xu et al. 2011).

Some authors proposed algorithms to perform the unsupervised point filtering/classification by studying the behavior of some statistical parameters of the LiDAR point cloud distribution values. Bartels et al. (2006) and Bartels and Wei (2006, 2010) have introduced a "*skewness* balancing" algorithm based on the analysis of elevations, that is able to distinguish ground and off-ground points, either in the case of flat and sloped terrains.

As well-known from Statistics, *skewness* (*sk*) and *kurtosis* (*ku*) are third- and fourth-order moments about the mean (see, e.g., Johnson et al. 1994). Sometimes they are grouped under the name of "higher-order moments". Their mathematical expression can be summarized as follows:

$$M^{(n)} = \left( \frac{1}{N\sigma^n} \sum_{i=1}^{N} (x_i - \mu)^n \right)^{1/n}, \quad (1)$$

where $n=3$ for skewness and $n=4$ for kurtosis, respectively; $N$ is the number of points; $x_i$ is the observed value of the $i$th point; $\mu$ and $\sigma^2$ are the mean value and the variance of population.

*Skewness* represents the degree of asymmetry around the mean. A skewness value of zero indicates a symmetric distribution. *Kurtosis* measures the degree of dominance of the distribution's peaks about its mean. Normal distribution has a $sk=0$ and $ku=3$. A normalized version of formula (1) can be found in the literature for kurtosis; in this case, normal distribution has $ku=0$. Figure 1 reports some examples of *probability distribution functions* (PDF) with skewness and kurtosis respectively smaller and larger than the ones of normal distribution.

Some distribution characteristics of the *elevation* values in LiDAR datasets can be highlighted by looking at both higher-order moments. In the case of skewness, negative values indicate dominance of "valleys", while positive values show dominance of "peaks". For kurtosis, values larger than 3 indicate a "peak" distribution, while values smaller than 3 characterize a "valley" distribution. In the case of *intensity* data, no physical interpretation of both parameters is meaningful.

As a preliminary assumption, a homogeneous class in a LiDAR dataset is expected to contain normally distributed values (Duda et al. 2001). Consequently, the analysis of both skewness and kurtosis can be useful for recognizing whether these parameters show a Gaussian behavior or not. In particular, this hypothesis should be satisfied for elevation data in the case of a flat area (sloped terrain entails some critical issues as discussed later) and for a homogeneous intensity data cluster. In Bao et al. (2007), the analysis of kurtosis point distribution is exploited for a classification into ground, buildings, and vegetation classes. A further improvement on the classification process was obtained by a combined analysis of both skewness and kurtosis for both elevation and intensity distribution values (Bao et al. 2009; Liu et al. 2009; Costantino and Angelini 2011).

In the mentioned literature, skewness and kurtosis are used within a classification scheme that is summarized in Fig. 2. Here the *intensity histogram* of a LiDAR dataset is considered at first (upper-right graph in

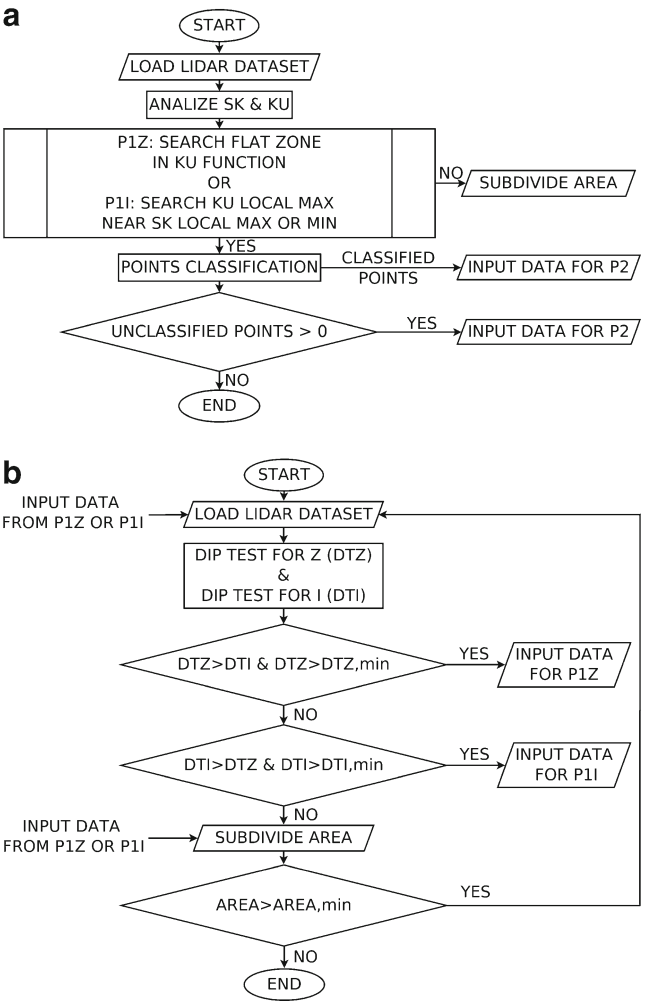



**Fig. 3** **a** Flowchart of the classification procedure for elevation ("*P1Z*") and intensity data ("*P1I*"), respectively. **b** Flowchart of the *bi/multi-modality* testing procedure "P2"

**Table 1** Error values in the four experiments; the lower row gives the error related to road extraction on the basis of intensity data from points that had been already classified as ground points

| | Exp. A1 (%) | Exp. A2 (%) | Exp. B1 (%) | Exp. B2 (%) |
|---|---|---|---|---|
| Total error | 5.60 | 2.69 | 8.90 | 1.40 |
| Type I error | 0.15 | 1.54 | 9.38 | 0.66 |
| Type II error | 20.79 | 6.60 | 7.05 | 1.80 |
| Error of road extraction | – | 31.66 | – | 20.19 |

Fig. 2). A preliminary analysis of the data variance, or another technique to check the presence of *bi-* or *multi-modality*, can be applied to avoid classifying a homogeneous dataset. In the example in Fig. 2, the presence of three trees in a field introduces a clear *bi-modality* in the dataset. By applying an iterative procedure, the point with the highest intensity value is sequentially removed from the dataset at each step. This operation results in discarding ground points that are characterized by the highest intensities (lighter points in Fig. 2). As soon as points are removed, both fractions of ground and tree points will become similar (second and third histograms from the right top in Fig. 2).

It is possible to notice how during the iterations (or *cycles*), when the ground intensity values are sequentially removed, skewness and kurtosis continuously change. When $sk=0$ and kurtosis has the minimum value, the distribution is symmetric and the same number of points is expected for the two different clusters. At this point, skewness and kurtosis start rising and the latter reaches a local maximum ($ku=3$) when nearly only vegetation points are present. This fact is also verified by a local maximum of skewness, confirming the presence of only one cluster made up of vegetation points. Because there is a range of intensity shared between both classes, Liu et al. (2009) suggested to stop the classification process at the iteration corresponding to the last local maximum of kurtosis. This point refers to the lower histogram in Fig. 2. All points with lower intensity are then collected into the same class ("trees" in this case). The same procedure can be applied again to the already removed points in order to seek for any possible subclasses. This behavior is true for intensity but not for elevation data. In this last case, the topography deeply affects both higher-order moments. As said before, negative skewness values indicate dominance of "valleys" while positive values show dominance of "peaks". Anyway, Liu et al. (2009) introduced an alternative approach to identify clusters of homogeneous points. In this case, the sequential classification procedure was applied by removing the points with the highest elevation values. Owing to the fact that these points belong to vegetation, once they have been removed, the two curves of skewness and kurtosis become and remain stable up to the end of process.

The examples reported in "Experiments" section will describe how the classification method based on skewness and kurtosis analysis can be exploited in a more effective procedure for classification or filtering of airborne LiDAR data. According to Vosselman and Maas (2010), the availability of multiple approaches is a fundamental prerequisite to accomplish this task in a successful way. This solution is important for both automatic (*unsupervised*) and interactive (*supervised*) procedures.

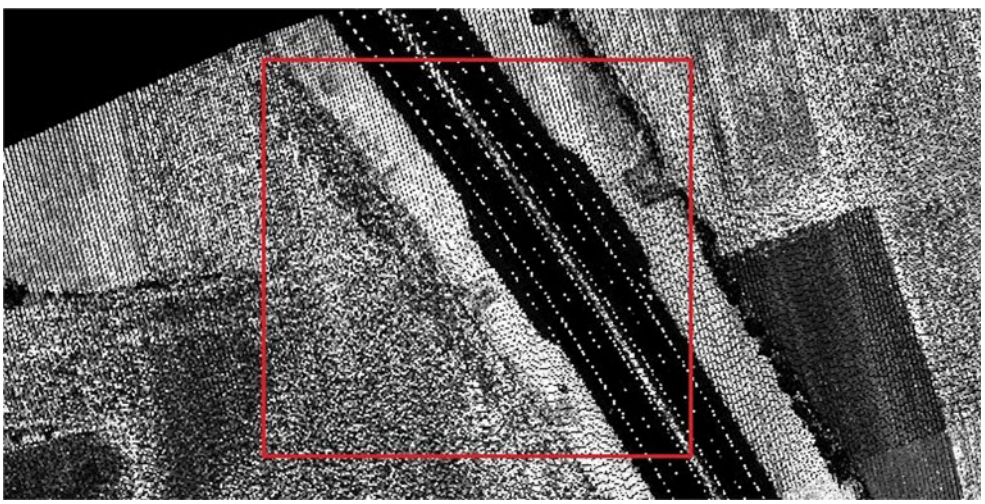


**Fig. 4** *Intensity image* of the area of experiment "A1"

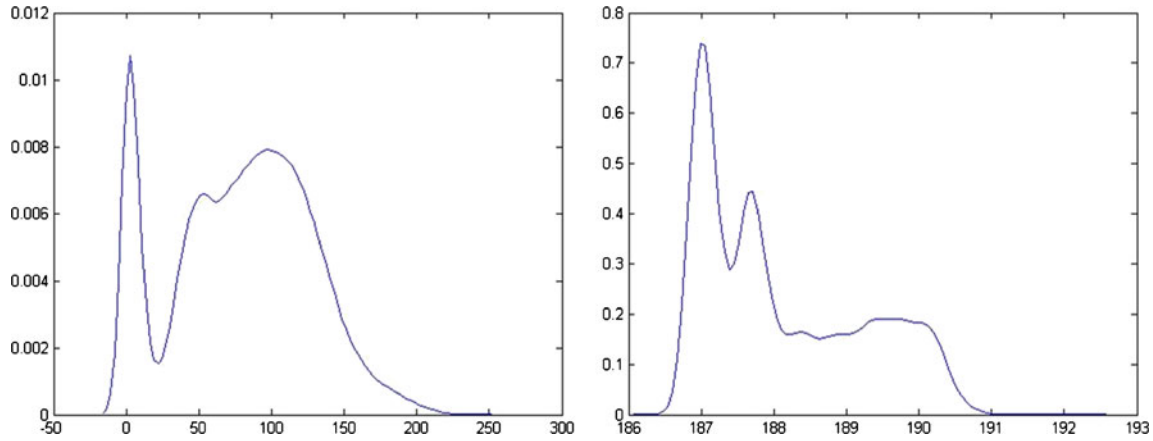



**Fig. 5** **a** Intensity distribution function and **b** elevation distribution function for the points from experiment "A1"

On the other hand, some aspects need to be improved for an extensive application of this method. The chance to work alternatively with both elevation and intensity data can be used for exploiting the property which is more likely to be successful. In the majority of the classification methods which rely on the higher-order statistical moments, their integration was not fully exploited. Real scenarios are not always as simple as the case studies reported in the most papers. Consequently, a strategy to decompose large scenes in smaller regions, where the basic discrimination criteria can be separately applied, is still pending. In addition, results achieved in different regions could be checked across one another to find discrepancies, to correct mistakes, and to merge those regions corresponding to the same category. Finally, some work is necessary to analyze how noise in the dataset might influence the classification process. Indeed, the same level of noise can affect more the classification of a smaller region than a larger one.

In the following sections, some solutions to cope more effectively with these concerns will be proposed and discussed. First of all, the description of the filtering/classification procedure is given in "The proposed filtering/classification procedure" section. "Experiments" section shows the application of the procedure to a few case studies coming from a real airborne LiDAR survey (see "Dataset description" subsection) with the purpose of validation. First, in "Analysis of simple scenes" subsection, the analysis of small scenes with an increasing level of complexity is carried out. Secondly, in "Analysis of complex scenes" subsection, larger and more involved regions are faced. Some final considerations on the achievements and the open problems are discussed in "Final considerations and future work" section, addressing to work to do in the future.

## The proposed filtering/classification procedure

Classifying and filtering LiDAR data based on *skewness* and *kurtosis* analysis is a proved methodology. However, as pointed out at the end of the "Introduction" section, some problems have to be overcome to setup

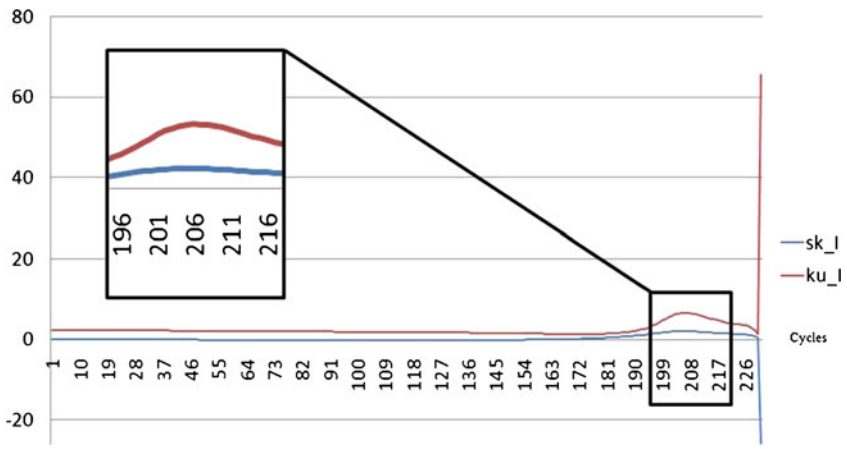



**Fig. 6** Skewness (*blue*) and kurtosis (*red*) behavior for the intensity values of experiment "A1"

an effective classification procedure capable of dealing with complex real scenes. Indeed, the complete automation of such procedure would be an important goal. The user would have only to check out the results and to label different classes. On the other hand, here the goal is to set up the basic algorithms and the general workflow of the filtering/classification method, while the user has still to make decisions. To better understand the processing pipeline, the case of a single homogeneous scene is considered at first. In a second stage, the extension to larger and more complex scenarios will be tackled.

The sequential procedure proposed by Liu et al. (2009) and described in the previous section is based on the alternative use of the most effective data category between elevation and intensity for classifying homogeneous point clusters. If the graph of the distribution is such to prefer the intensity values, since it shows a clear *bi-* or *multi-modality*, from the analysis of skewness and kurtosis, the last part of the distribution values will be classified as in Fig. 2. Points relating to the right side of the kurtosis local maximum are assigned to the same class, while points relative to its left part are not classified. A similar approach is still valid if the point classification is carried out on the basis of elevations. Points that satisfy a local flat condition for the last part of skewness and kurtosis functions are homogeneously classified, while the others remain unclassified. The same procedure can be applied again to all the points, but using the complementary data category, the intensity analysis is applied to the data already classified on the basis of elevations and vice versa. This mixed procedure might identify further subclasses within the already classified points and to classify some points that are still unclassified.

The procedure starts with a given point cluster and the first step is to check whether it corresponds to a unique class. This step is accomplished by fitting both elevation and intensity data with a plane. The analysis of the residuals with respect to the estimated plane allows one to assign all points to a single class if the root mean square (RMS) of residuals is less than a predefined threshold. In this case, the classification procedure is halted, otherwise the next processing stages can be tackled.

If the process goes on, the following step is the analysis of the point elevation and intensity *histograms* to select the data category to be considered at first. Owing to the irregularity of histograms, a non-parametric estimation of the PDF which better fits each histogram is applied. Here the approach proposed in Epanechnikov (1969) was followed. A PDF can be obtained by a convolution process of a chosen kernel applied to each sampled value. Given a data set ($x_1$, $x_2$, …, $x_n$) sampled from a distribution having an unknown PDF $f(x)$, the problem is to estimate the shape of this function from the following relationships:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i), \quad \text{with } K_h = \frac{1}{h}K\left(\frac{x}{h}\right), \tag{2}$$

where $K(x/h)$ is the kernel, i.e., a non-negative density function with integral equal to 1; $h>0$ is a real positive parameter defining the size of the sampling class (the default value is

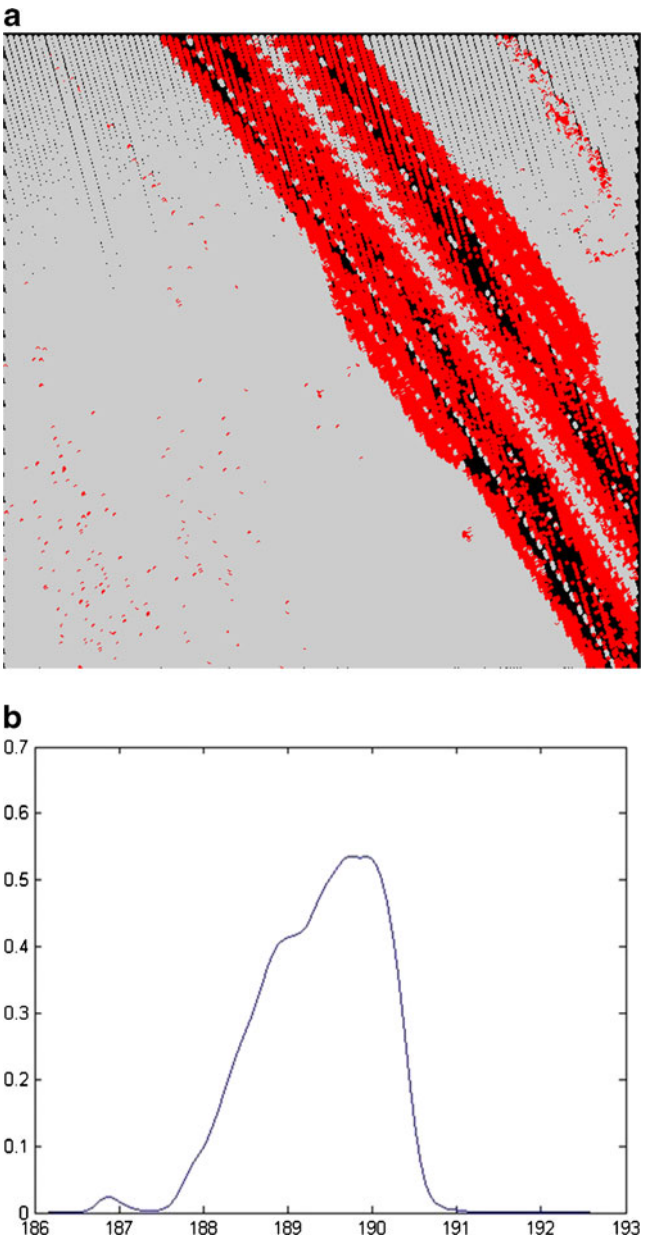



**Fig. 7** **a** Point classification based on intensity data (experiment "A1"); classified points at this stage are depicted in red. **b** PDF for the elevation values of the classified points

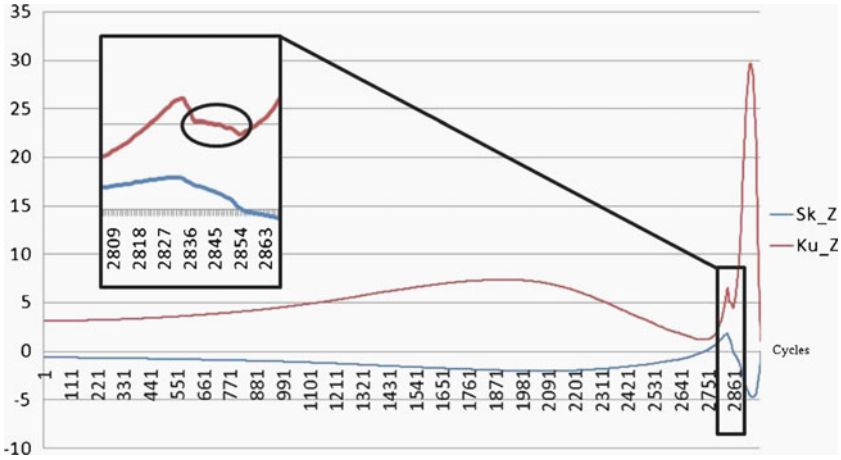



**Fig. 8** Skewness (*blue*) and kurtosis (*red*) values for elevation of the points classified in *red* in Fig. 7a (experiment "A1")

100). Symmetrical density functions, with respect to the origin, are usually applied (a normal PDF in this case).

The selection of the data category to start with falls on the one that better shows a *bi-* or *multi-modality*. If the clusters are totally disjoint, the problem is trivial. If this is not the case, further analyses are carried out. Hartigan and Hartigan (1985) proposed to apply the "dip test" to measure *multi-modality* in a sample by the maximum difference, over all sampled points, between the empirical distribution function and the *unimodal* distribution function that minimizes that maximum difference. More recently, *profile analysis* was carried out by applying different strategies like the *Bayesian Information Criterion* score (see, e.g., Yeoung et al. 2001) and the *variational Bayesian* approach (see, e.g., Teschendorff et al. 2005). Teschendorff et al. (2006) proposed integrating the analysis of kurtosis into the previous models. They showed that in the case of a *bi-modal* distribution, a mixture of two approximately equal mass Gaussians must have $ku<3$, whereas $ku>3$ in the case of highly unequal masses. They also found out a relationship among kurtosis, the standardized separation between the two clusters and the minor cluster mass (as fraction of the total). Practically, comparing two distribution functions, the best seems to be the one with the *kurtosis* value much lower than 3. However, here the "dip test" is applied to select between elevation and intensity data (see the flowchart of the process "P2" in Fig. 3b). In addition, if no significant difference is found by this test, the user is asked to make a decision about the data property to use or to stop the classification process.

In the case the selection of elevation and intensity values is positively accomplished, the skewness and kurtosis variation analysis is applied to identify a significant point cluster through the point removal cycle. Afterwards, the selected cluster is analyzed by considering the complementary data category. Then the procedure is repeated for the remaining unclassified data. Figure 3a shows the flowchart of the procedure "P1" for the elevation ("P1Z") and for intensity data ("P1I"), respectively.

The classification method here proposed works well for small areas, where the presence of only a few *modal* distribution values can be expected for elevation and intensity. The method becomes prohibitive when applied to large, not homogeneous, and complex areas, where both data categories might show a *multi-modal* behavior.

This is the reason why the classification procedure "P2" was thought as a "progressive multi-analysis method", where the whole area is subdivided into regular sub-areas to be independently processed. Some first experimental results confirmed the suitability of the filtering/classification method to cope effectively with

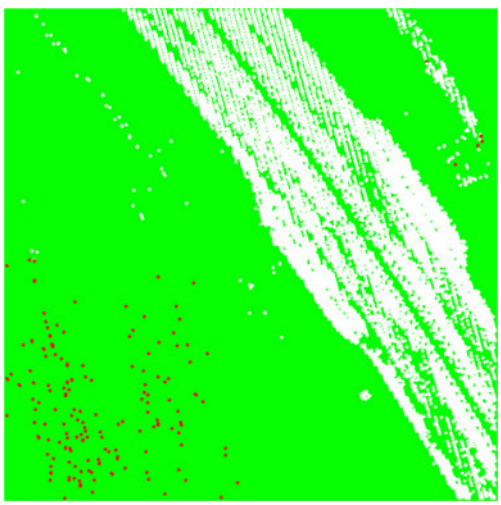


**Fig. 9** Red points are classified on the basis of elevation, after the first classification based on intensity data (see Fig. 7a—experiment "A1"); points in *green* are not considered

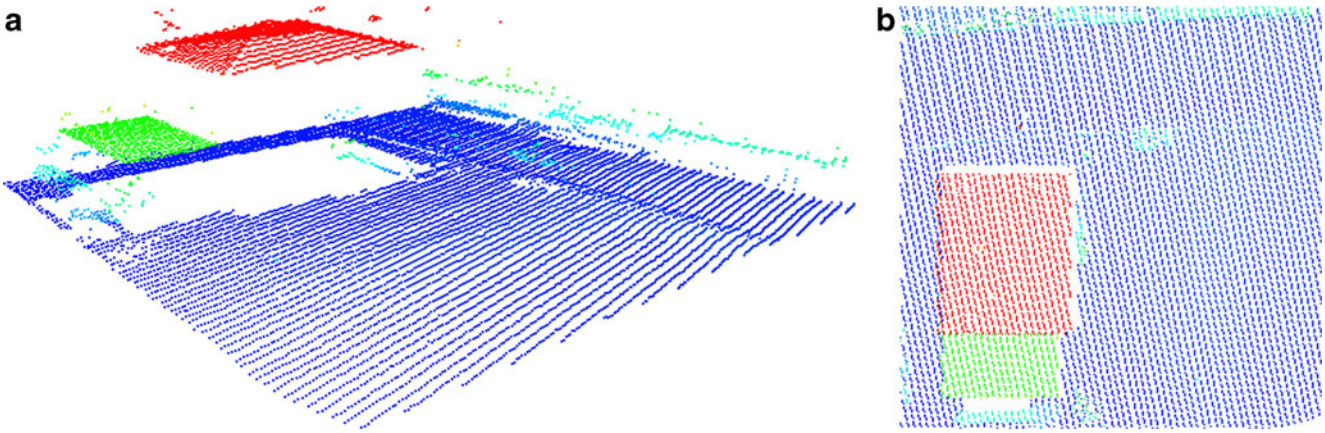



**Fig. 10** 3D view (on the *left*) and 2D view (on the *right*) of the area of experiment "A2;" colors change according to elevation values from *blue to red*

complex scenes (see "Analysis of complex scenes" subsection). As can be seen in Fig. 3b, the workflow of procedure "P2" is extended when working with complex and large regions.

In the current implementation, the dataset has been inscribed into a rectangular window. "No-data" values are assigned where there are no LiDAR data. If the classification procedure cannot be successfully applied to the whole region, this is split up into regular rectangular sub-windows. Then the process is applied independently to each of them. Only those sub-windows that cannot be classified are sub-divided again. The procedure is repeated as far as all regions will be classified. An upper limit to the sub-window levels is established to avoid processing of too smaller regions. A reasonable alternative to the subdivision of large datasets into smaller regular sub-regions is to apply a preliminary rough classification based on a well-assessed state-of-art technique. Moreover, the strategy for splitting the larger datasets do not influence the basic classification algorithm, which does not require any data regularity and can process sparse point clouds as well.

Once this process is accomplished, each region will be independently classified. At the moment, the user tries to identify classes shared by different regions. This task is simply carried out by assigning the same label to the classes corresponding to the same category of objects. In "Dataset description" subsection more details about this aspect will be given.

The establishment of a common classification framework gives also the chance to check inconsistencies and mistakes left by the independent classification in each region, for instance by exploiting the topological properties of neighboring classes as suggested in Forlani et al. (2006).

At the current state of this research, different tasks have not been implemented yet in an automatic way. Indeed, hitherto the aim has been to establish and assess some criteria to support the decision making process in a "supervised" manner.

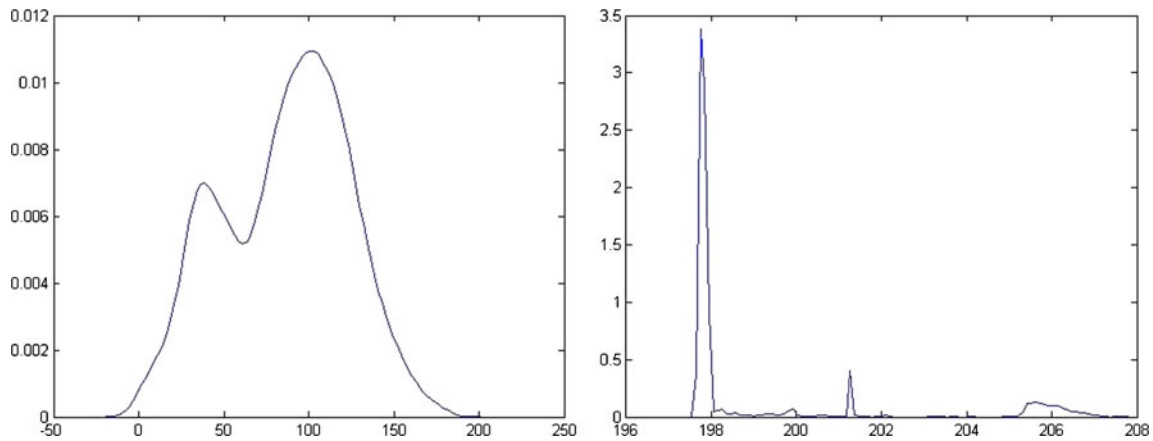



**Fig. 11** **a** Intensity and **b** elevation PDF for the points of experiment "A2"

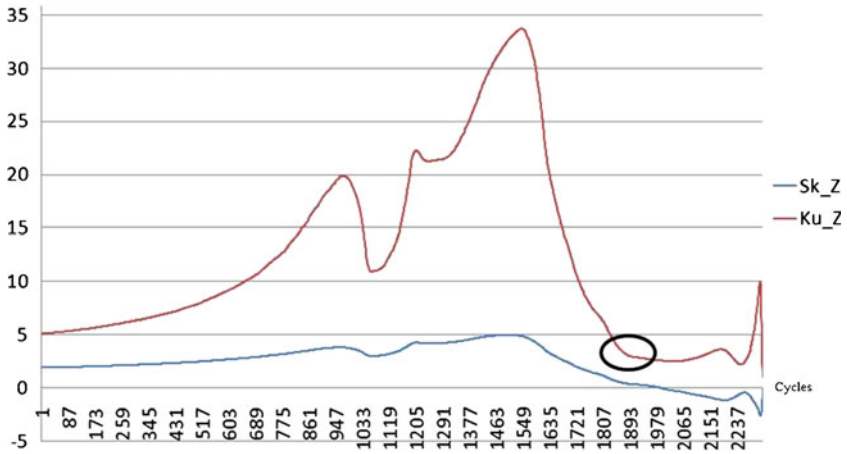



**Fig. 12** Skewness (*blue*) and kurtosis (*red*) plots for the point elevation values of experiment "A2"

## Experiments

The performances of the sequential classification procedure have been verified by some experiments. Some regions extracted from the LiDAR dataset described in "Dataset description" subsection have been utilized to this purpose.

Experiments have concerned two different kinds of scenes: at first, simple, small scenes have been focused on ("Analysis of simple scenes" subsection); then two larger scenes have been afforded ("Analysis of complex scenes" subsection).

The performance of the algorithm is measured by comparing the obtained classification against a manual classification of the same reference data. Table 1 shows the total error (number of misclassified points as a fraction of all points), the type I error (number of misclassified ground points as a fraction of all ground points), and type II error (number of misclassified points as a fraction of all points in the selected class).

### Dataset description

Data used for experiments are relative to two strips of a LiDAR survey over the municipality of Tavagnacco (Udine, Italy). The flight was carried out in 2007 with a Leica ALS50 sensor. Forty strips, covering an area of more than 30 km$^2$, for more than 400 millions of points, have been acquired at a flight height of 1,000 m a.s.l. with a point density of ca. 13 points/m$^2$. The mean linear distance among the surveyed points is 28 cm. The available (not re-sampled) data concern the first-echo response only. Consequently, classification did not rely on multiple responses, although this information could help, especially for the classification of vegetation.

A manual point classification had been previously carried out by the program MARS Explorer 6.1 (Merrick and Company 2010). Four target classes have been manually identified: "ground", "streets", "buildings", and "vegetation and other objects" (i.e., cars).

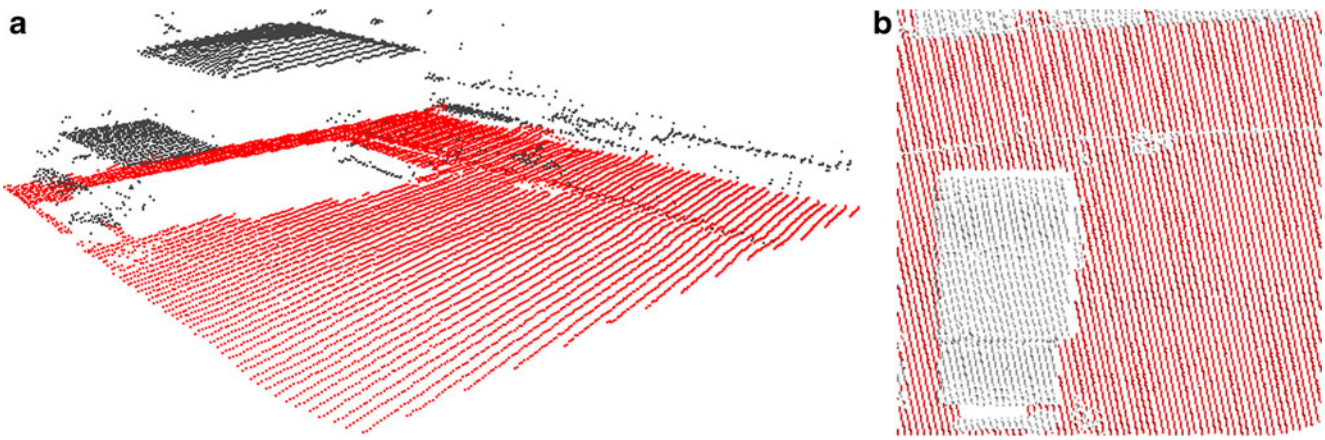



**Fig. 13** Classification of points into ground (*red points*) and off-ground points (*black points*) based on elevation data for experiment "A2" (3D view on the *left* and 2D view on the *right*)

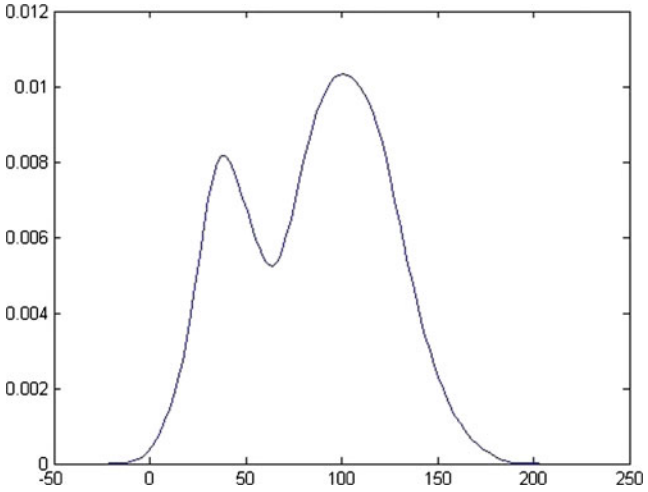



**Fig. 14** PDF of the intensity values for the ground points of experiment "A2"

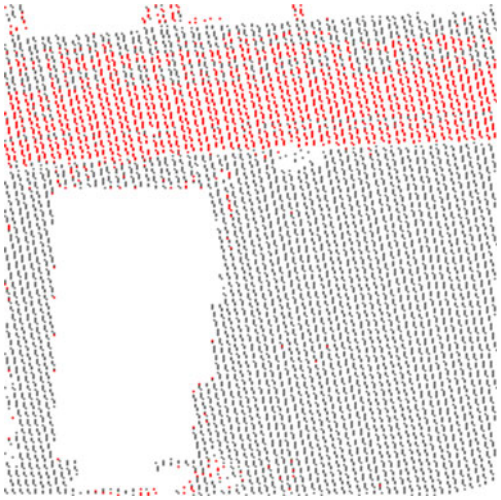


**Fig. 16** Classification based on intensity of the ground points of experiment "A2"

The applied filtering/classification method in the adopted software provides the analysis of individual sections along the strip with the option "Place Adjustable Profile Line". This option allows the operator to create a section with an input thickness. Afterwards, the option "Edit Toolbars" has been applied in order to label a class of selected points.

At the end of the manual classification process 11,932,445 points have been classified. Out of these, 80.4 % of the total points have been classified as "ground", 8 % as "vegetation and other objects", 8.2 % as "buildings," and 3.4 % as "streets."

## Analysis of simple scenes

### *Experiment "A1": road extraction*

The experiment is related to one of the main applications of the laser point analysis, i.e., *road extraction*. To this aim, a small area of the dataset, crossed by a highway, is taken into account (Fig. 4).

By analyzing the distribution functions obtained from the histograms for both intensity (Fig. 5a) and elevation data (Fig. 5b), the former shows a clear *bi-modality* for what concerns the intensity. On the contrary, it is hard to distinguish any point cluster in the elevation data. This is confirmed by the values of the "dip test" that furnishes 0.0166 for elevation and 0.0543 for intensity. Thus, the choice falls on the computation of skewness and kurtosis variations along the cycles for the intensity values. Throughout the paper, the graph of skewness behavior is represented with a blue line, while that of kurtosis with a red line, as in Fig. 6.

The decision is to classify the points in correspondence of the maximum value of kurtosis, where a "peak" value of *skewness* is also present at cycle 206. The result is shown in Fig. 7a, where classified points are represented as red dots and unclassified points as gray ones.

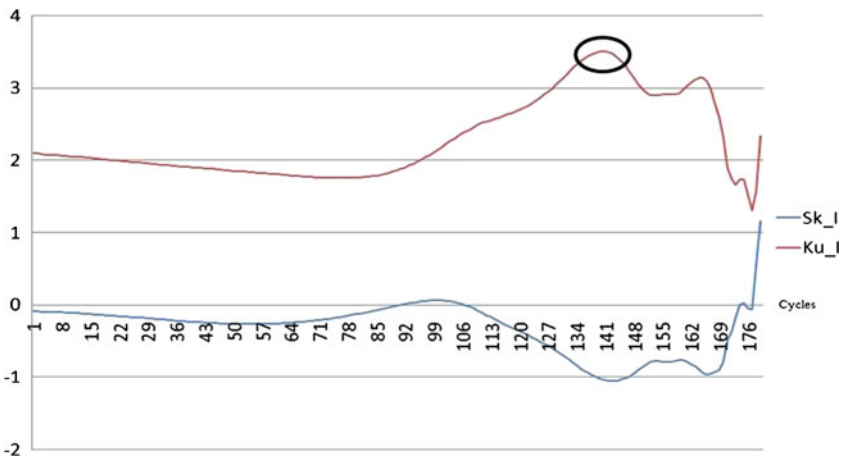



**Fig. 15** Behavior of skewness (*blue*) and kurtosis (*red*) for the intensity values of the classified ground points (experiment "A2")

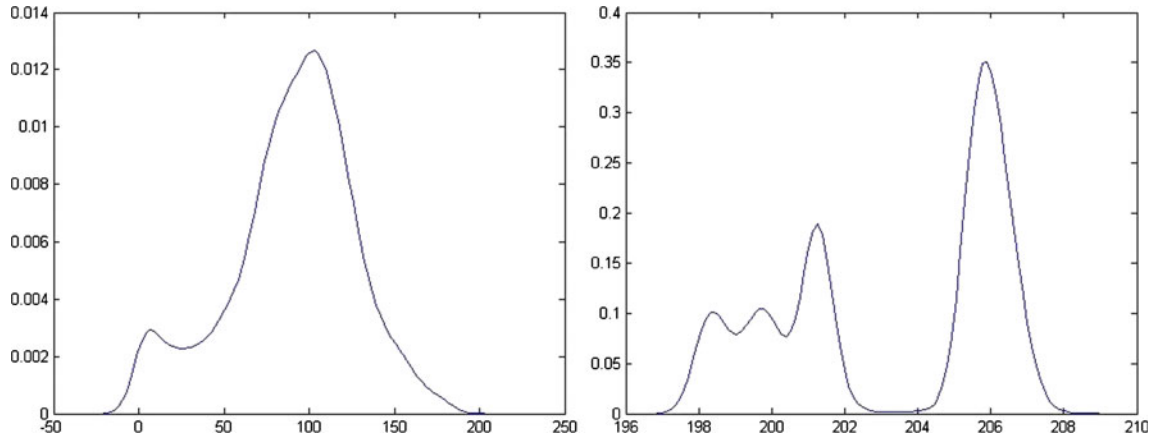



**Fig. 17** **a** Intensity and **b** elevation PDF for the unclassified points of experiment "A2"

As expected, the largest part of the points belonging to the asphalted area is correctly classified, including some points belonging to vegetation (upper right part of Fig. 7a) and some sparse ground points. A provisional classification label is assigned to these points, while the ones not yet classified (gray points in Fig. 7a) are considered as still unclassified.

According to the flowchart in Fig. 3a, b, the elevation analysis is then applied to the classified data (i.e., red points in Fig. 7a). As can be seen in the PDF diagram shown in Fig. 7b, points belonging to the small cluster at height 187 m should be separated from the ones in the range 188–191 m. After applying again the skewness and kurtosis variation analysis for the elevation values of such points, the behavior shown in Fig. 8 is obtained.

On the basis of elevation analysis, the local flatness of kurtosis suggests the classification of points at cycle 2,845. The result makes it possible to extract from the red points in Fig. 7a those belonging to the "ground" class (Fig. 9).

As can be seen in Table 1, a quite large type II error was found, showing in this case the limit of the proposed method for extracting road points.

*Experiment "A2": extraction of multiple classes*

This experiment represents a very significant synthesis of common real situations (see Fig. 10), with the presence of ground, vegetation, one road, and one building.

The PDF for the intensity and elevation data are shown in Fig. 11. Here the shape of PDF plot for elevations reveals more than two classes, while the PDF of intensities mainly indicates two partially overlapped classes. According to the result of the "dip test", the classification process started from the elevation values (Fig. 11).

Points are classified in correspondence of cycle 1,890, where the kurtosis plot features a local flat area (see Fig. 12). This first classification splits points into ground and off-ground classes (Fig. 13). Then, the ground points are classified again on the basis of the intensity data. Figure 14 reports the behavior of the ground point distribution analyzed by using intensity. After computing skewness and kurtosis indices, the plots reported in Fig. 15 are obtained. Points are classified according to cycle 141, in correspondence of a local maximum of kurtosis. In this way, it is

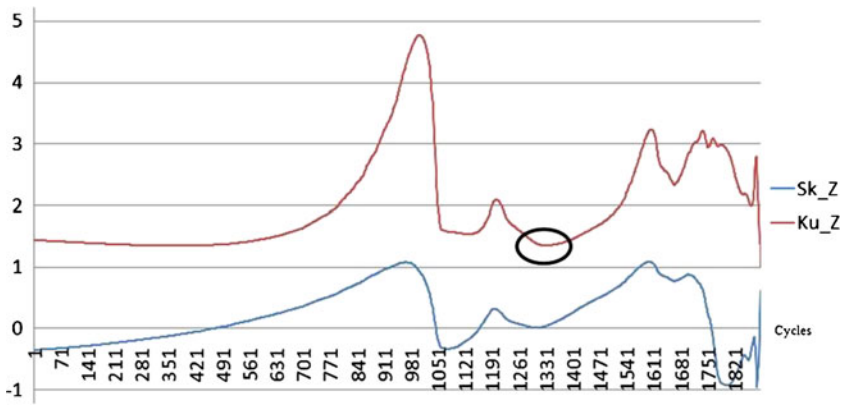



**Fig. 18** Plots of skewness (*blue*) and kurtosis (*red*) for the elevation of the unclassified points (experiment "A2")

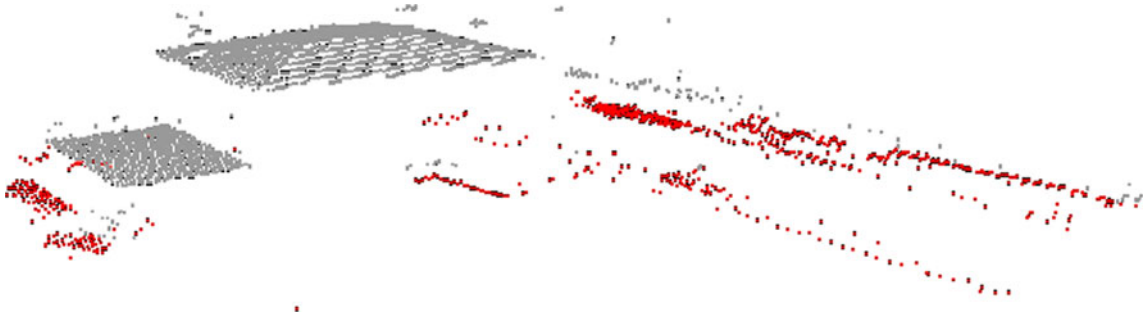


**Fig. 19** Segmentation based on elevation of vegetation points (*red*) and building points (*gray*) from experiment "A2"

possible to separate those points belonging to road surface (see Fig. 16).

Points not yet classified are now taken into account and their distribution evaluated. The PDF of elevations (Fig. 17b) shows more than two classes, while intensity data are basically unimodal. Consequently, the classification process carries on by considering the elevation values (Fig. 18). Points are classified according to the elevation following cycle 1,319,

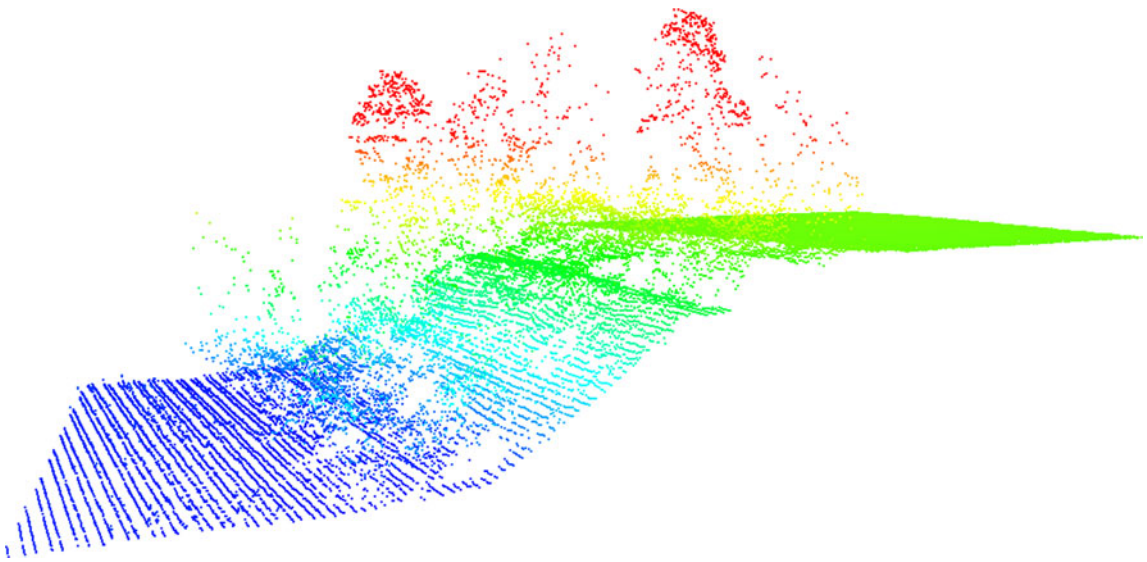


**Fig. 20** 3D view of the "B1" experiment area colored according to elevation from blue to red

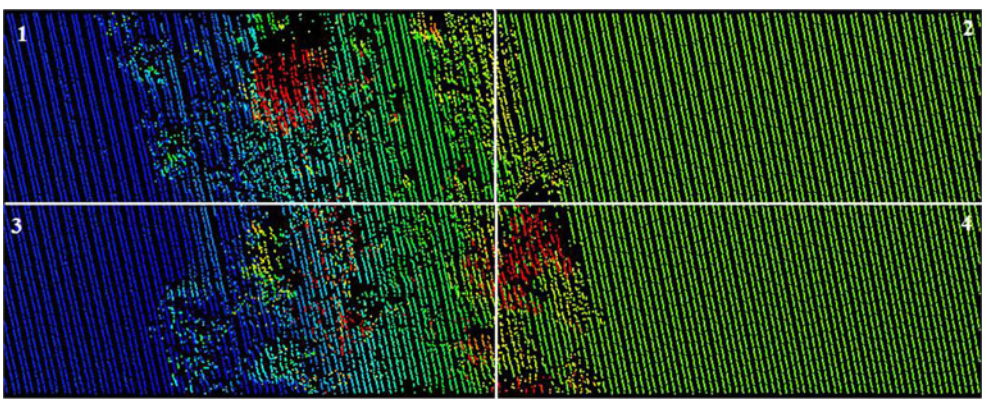



**Fig. 21** First subdivision of the entire area of experiment "B1" in four parts

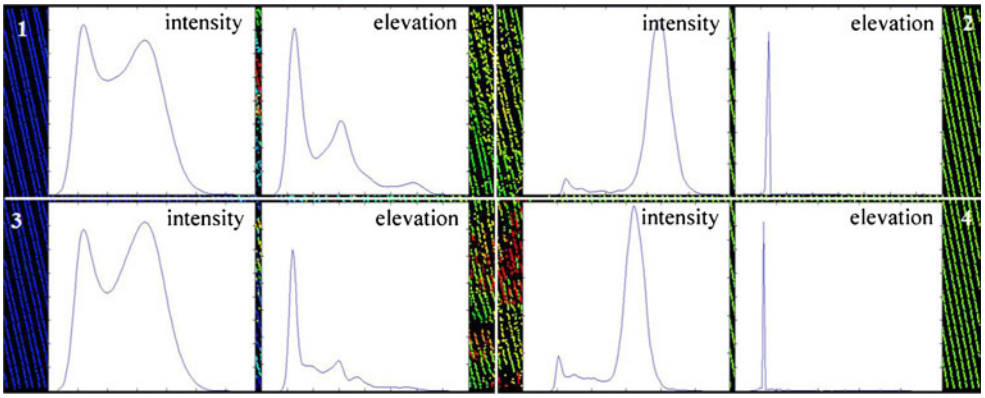



**Fig. 22** PDF for intensity and elevation data per each of the four main areas (experiment "B1")

in correspondence of a local flatness of kurtosis. The result made possible to successfully split up vegetation and building points (Fig. 19).

The results of comparison with benchmarking data in Table 1 look quite satisfying for this dataset. In this case, type II error relates the correct classification of buildings.

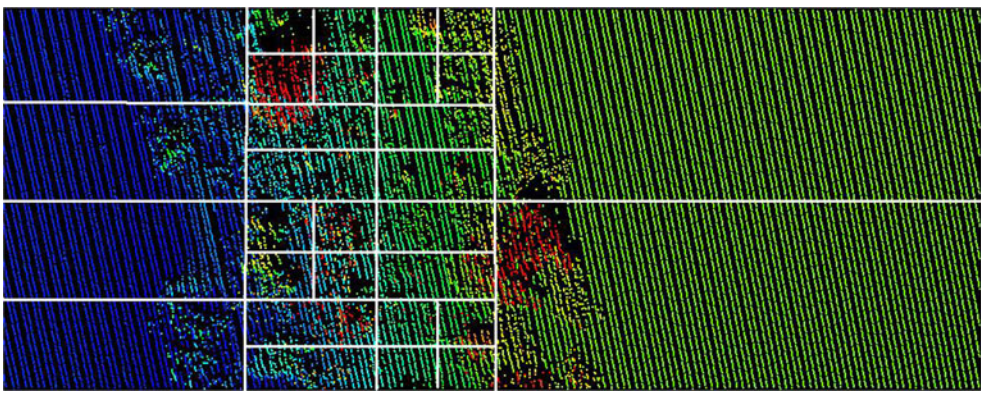


**Fig. 23** Final subdivision of the entire area of experiment "B1"

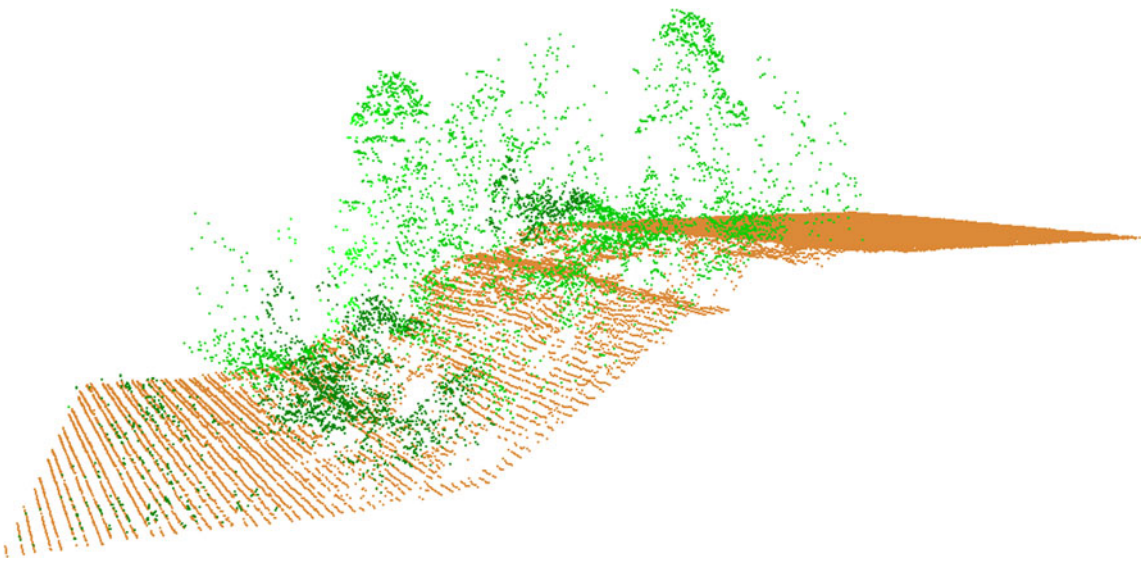


**Fig. 24** 3D view of the final classification result of experiment "B1": "ground" class in brown, "vegetation and other objects" class in *green*

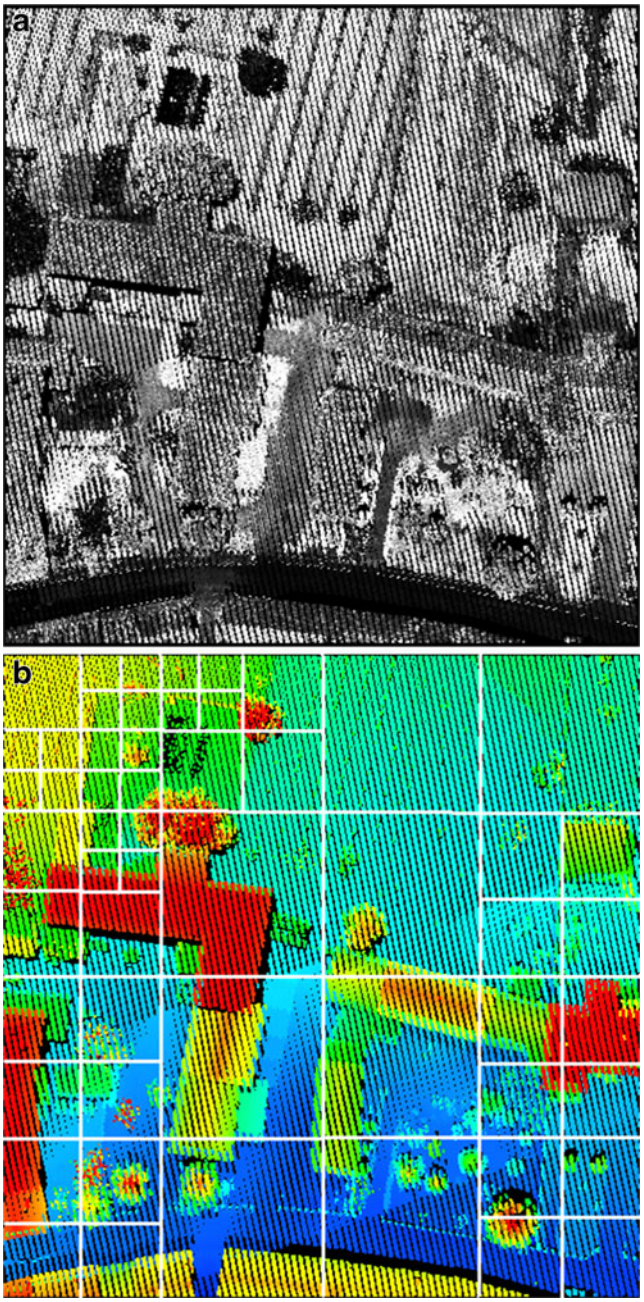



**Fig. 25** Intensity (*above*) and elevation (*below*) image of the area of experiment "B2"

The only exception is the error on road extraction based on the analysis of intensity from points previously classified as "ground" (31.66 %).

Analysis of complex scenes

It was already mentioned that in the case of complex scenes, it is necessary to subdivide the whole area in sub-regions where the classification procedure can be successfully applied. In the following, the results of two experiments in such areas are reported. The first test relates to an area covered by vegetation, characterized by two flat and one sloped parts. The second experiment refers to a complex urban environment, characterized by various kinds of features.

*Experiment "B1": vegetation extraction*

The experiment was carried out for the complex area reported in Fig. 20. This area is characterized by two flat parts, located at different height, connected by a sloped terrain covered by trees and other kinds of vegetation.

At the beginning the whole area is subdivided into four zones (Fig. 21) and for each zone the intensity and the elevation distribution values are computed (Fig. 22). According to the distribution results of the zones 1 and 3, it is decided to still divide these zones in four sub-parts. In this case, the use of preliminary rough classification is also expected to work well, but it has not been applied here.

By proceeding this way, after the analysis of the shape of PDF for elevation and intensity data, the whole area is finally subdivided according to the layout reported in Fig. 23. The skewness and kurtosis analyses for both data categories of each zone result in the classification shown in Fig. 24, where points classified as "ground" are depicted in brown and those as "vegetation and other objects" in green.

The average value of three types of errors in Table 1 is ca. 7–9 %, result that is fairly good if compared to other state-of-art methods (Vosselman and Maas 2010). Type II error here means the correct classification of vegetation points.

*Experiment "B2": road and terrain extraction*

The area (10,000 $m^2$) is characterized by a complex urban environment (Fig. 25) and the density of surveyed points is higher than in other datasets due to the overlap between two strips (26.5 points/$m^2$). In the upper-left part, there is a sloped terrain with presence of a building and vegetation. The other areas are characterized by the presence of buildings, roads, complex vegetation, and flat terrain.

The whole area is subdivided into small sub-regions (Fig. 25, image at bottom) and for each zone the intensity and the elevation distribution values are computed. By proceeding as in the previous experiment, the result of the classification is reported in Fig. 26.

The filtering total error results equal to 1.40 %. Type I error corresponds to 0.66 %, while type II error, i.e., the classification error of "vegetation and other objects" as well as "buildings" classes, is equal to 1.80 % (see Table 1).

In this experiment, from classified ground points, the road points were extracted by using intensity with the

iterative removal method (Fig. 26, right image) and the error (i.e., number of misclassified road points as a fraction of all road points) result is equal to 20.19 %. The same type error was found in the experiment "A2" (31.66 %).

Discussion

According to the reported results, the algorithm seems to work well for filtering ground points, also in the case of heavily vegetated slopes. In fact, as reported in Sithole and Vosselman (2005), such kind of terrain is very often not correctly classified with standard filtering methods. Moreover, the type II error value would be significantly reduced in case of buildings and other kinds of man-made objects.

In the case of complex scenes, the subdivision of the full dataset in rectangular areas worked effectively in the reported examples. Other principles, like preliminary pre-classification in free-form polygonal regions which feature homogeneous properties, could be exploited as well.

A further problem that might affect precision and reliability of the results is the influence of different kinds of noise in the dataset. In particular, it would be really important to analyze the behavior of skewness and kurtosis variations with respect to data errors. Indeed, the adopted method selects a class of data at the end of an iterative removal process. This means that the computation of higher-order moments for the last data characterizing a particular class might be strongly influenced by the presence of outliers. For this reason, skewness and kurtosis robust measures should be considered in future instead of the traditional ones (e.g., Mardia 1974; Kim and White 2004).

## Final considerations and future work

The paper proposes a new LiDAR point filtering and classification method based on the sequential skewness and kurtosis analysis of elevation and intensity point distribution values. This technique introduces some improvements with respect to other similar approaches reported in the literature. Indeed, the basic algorithm here implemented works well for simple scenes, after removing at each step of the process the largest data values as suggested by Liu et al. (2009). First of all, some solutions have been included to better support user's decisions. Furthermore, a procedure to split larger datasets into smaller regions to be independently processed has been added here to extend the application to complex scenes.

Some numerical experiments with real airborne LiDAR data confirmed that classification errors (both I and II type errors) are comparable and, in some cases, also better than the ones reported in the literature when using other well assessed techniques. Moreover, the applicability of the method has been also demonstrated for filtering out off-ground points in the case of vegetated slopes.

In the current implementation, the procedure runs as an interactive procedure, where the user makes decisions along the classification process. After further developments, this method is prone to become an *unsupervised* classification method. On the other hand, the user knowledge still will

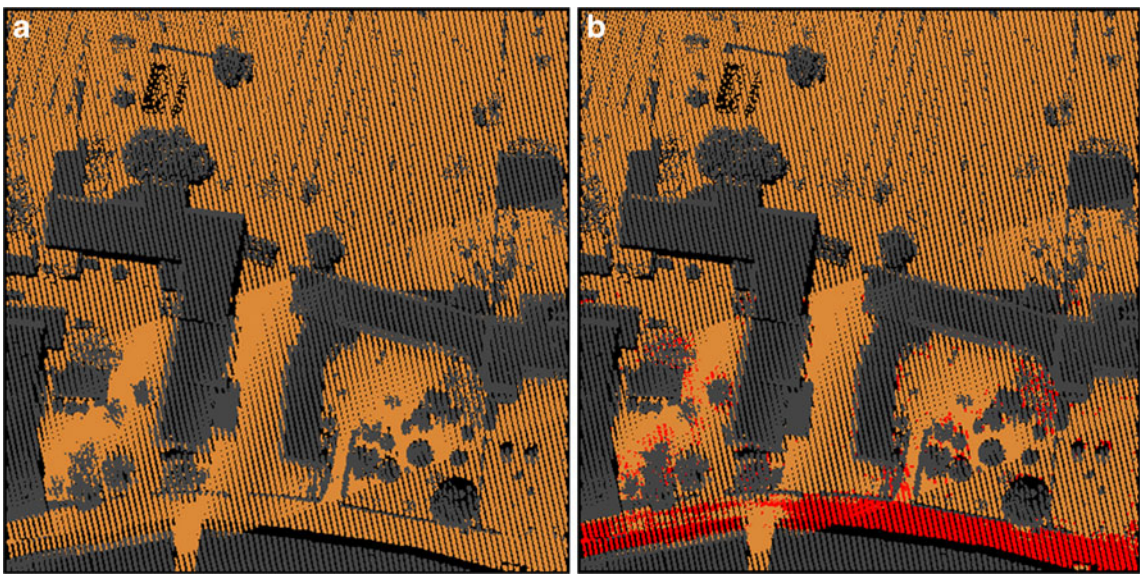



**Fig. 26** On the left, all points classified as "terrain" are shown (*brown points*); from these points, the ones belonging to the "road" class (*red points* in the *right sub-figure*) are extracted (experiment "B2")

remain fundamental for the discrimination of critical cases, as in the most classification procedures.

Finally, the evaluation of noise influence is a crucial point. The core classification algorithm runs on the basis of iterative elimination of points from a dataset. This means that mistakes and noise can have a largely different influence in the first and last cycles, owing to the decreasing number of remaining points. This effect is particularly critical because the classification is run on the points left at the end of each cycle. Procedures to make the process more robust will be investigated in the future.

**Acknowledgments** The authors thank Prof. Andrea Fusiello for proofreading the paper and Dr. Daniele Piccolo for providing some statistical testing computations by the R package. Acknowledgements go to Kwang-Hua Fundation (Tongji University, Shanghai, People's Republic of China). This research was partially funded by the National High-tech R&D Program of China (no. 2012AA121302), and by the National Basic Research Program of China (no. 2013CB733204).

## References

Axelsson P (2000) DEM extraction from laser scanner data using adaptive TIN models. Int Arch Photogramm Remote Sens 33(B4):110–117

Bao Y, Cao C, Chang C, Li X, Chen E, Li Z (2007) Segmentation to the clouds of LIDAR data based on change of Kurtosis. Proc of SPIE. doi:10.1117/12.791521, International Symposium on Photoelectronic Detection and Imaging 2007: Image Processing, edited by Liwei Zhou, Vol. 6623, 66231N

Bao Y, Cao C, Zhang H, Chen E, He Q, Huang H (2009) Synchronous estimation of DTM and fractional vegetation cover in forested area from airborne LIDAR height and intensity data. Sci China Ser E 51(2):176–187

Bartels M, Wei H (2006) Segmentation of LIDAR Data using measures of distribution. Int Arch Photogramm Remote Sens Spat Inf Sci 36(7)

Bartels M, Wei H (2010) Threshold-free object and ground point separation in LIDAR data. Pattern Recogn Lett 31:1089–1099

Bartels M, Wei H, Mason D (2006) DTM Generation from LIDAR Data using Skewness Balancing. In: Proc. of '18th Int. Conference on Pattern Recognition (ICPR'06),' Vol. 1, Hong Kong, China S.A.R., 20–24 Aug 2006, pp. 566–569

Costantino D, Angelini MG (2011) Features and ground automatic extraction from airborne LiDAR data. Int Arch Photogramm Remote Sens Spat Inf Sci 38(5/W12)

Crosilla F, Macorig D, Sebastianutti I, Visintini D (2011) Points classification by a sequential higher—order moments statistical analysis of LIDAR data. Int Arch Photogramm Remote Sens Spat Inf Sci 38(5/W12)

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley & Sons, New York, USA

Epanechnikov VA (1969) Non-parametric estimation of a multivariate probability density. Theor Probab Appl 14:153–158

Forlani G, Nardinocchi C, Scaioni M, Zingaretti P (2006) Complete classification of raw LIDAR data and 3D reconstruction of buildings. Pattern Anal Appl 8(4):357–374. doi:10.1007/s10044-005-0018-2

Haralick RM, Shapiro LG (1992) Computer and robot vision. Addison-Wesley, Boston

Hartigan JA, Hartigan PM (1985) The dip test of unimodality. Ann Stat 13(1):70–84

Johnson NL, Kotz S, Balakrishan N (1994) Continuous univariate distributions, vol 1, 2nd edn. Wiley, New York

Kim TH, White H (2004) On more robust estimation of skewness and kurtosis. Financ Res Lett 1(1):56–73

Kraus K, Pfeifer N (1998) Derivation of digital terrain models in wooded areas. ISPRS J Photogramm 53(4):193–203

Liu Y, Li Z, Hayward R, Walker R, Jin H (2009) Classification of airborne LIDAR intensity data using statistical analysis and Hough transform with application to power line corridors. In: Proc. '2009 Digital Image Computing: Techniques and Applications,' Melbourne, Australia, 1–3 Dec 2009, pp. 462–467

Mardia KV (1974) Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya Ser B 36:115–128

Merrick & Company (2010) http://www.merrick.com/index.php/geospatial/services-gss/mars-software. Accessed 24 March 2013

Sithole G, Vosselman G (2004) Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds. ISPRS J Photogramm 59:85–101

Sithole G, Vosselman G (2005) Filtering of airborne laser scanner data based on segmented point clouds. Int Arch Photogramm Remote Sens Spat Inf Sci 36(3/W19):66–71

Teschendorff AE, Wang Y, Barbosa-Morais N, Brenton JD, Caldas C (2005) A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. Bioinformatics 21:3025–3033

Teschendorff AE, Naderi A, Barbosa-Morais N, Caldas C (2006) Pack: profile analysis using clustering and kurtosis to find molecular classifiers in cancer. Bioinformatics 22:2269–2275

Vosselman G, Maas HG (2010) Airborne and Terrestrial laser scanning. Whittles Publishing, Dunbeath

Xu Z, Liu L, Liu X (2011) A new filtering algorithm for LIDAR data fused with image segmentation information. SPIE 8286, paper no. 828610

Yeoung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17:977–987