A SEGMENTATION PROCEDURE OF LIDAR DATA BY APPLYING MIXED PARAMETRIC AND NONPARAMETRIC MODELS

Fabio Crosilla, Domenico Visintini, Francesco Sepic

Department of Georesources & Territory, University of Udine, via Cotonificio, 114 I-33100 Udine, Italy
crosilla@dgt.uniud.it

KEY WORDS: LIDaR data, segmentation process, parametric and nonparametric models

ABSTRACT:
The paper proposes a segmentation procedure inspired to a robust LIDaR filtering data method recently introduced by the authors. The method is based on the application of a Simultaneous AutoRegressive (SAR) model for describing a trend surface and of an iterative Forward Search (FS) algorithm to detect clusters of non-stationary data.
The procedure consists in an automatic process to identify raw clusters of data relating to the geometrical configurations to be segmented with the robust iterative SAR-FS parametric model. The search of homogenous clusters of points is carried out by applying a local polynomial regression algorithm, automatically adapted to the morphological variability of the LIDaR points.
The combination of the parametric and nonparametric models in a mixed analytical procedure makes it possible to optimize the efficiency of the segmentation and dramatically reduce the requirements of computational memory and time consuming.
Some significant experiments make it possible to evidence the potential of the method proposed.

1. INTRODUCTION

Airborne Laser Scanning technique is extremely efficient to fulfil increasing demand of high accuracy spatial data for civil engineering, environmental protection, planning purposes, etc.. The main processing steps are the filtering of the points (to detect the ground terrain), their segmentation (to classify the point dataset in different classes), and the 3D modeling of clusters (to enhance the data structure from an irregularly sparse to a vector object-oriented one). With regard to point segmentation algorithms, exploiting geometrical and/or radiometric properties, this paper proposes a new one inspired to the procedure suggested by Crosilla, Visintini and Prearo (2004a) for the filtering of non-ground measurements from airborne laser data. The method is based on a Simultaneous AutoRegressive (SAR) model to describe the geometrical trend of a surface (chapter 2), and on an iterative Forward Search (FS) algorithm (Atkinson and Riani, 2000), to find out outliers and/or clusters of non-stationary data (chapter 3). Starting from a subset of stationary LIDaR data, the forward search approach allows to perform a robust iterative estimation of the SAR unknown parameters. At each iteration, one or more LIDaR points are joined, according to their level of agreement with the postulated surface model. Outliers and or non-stationary data are identified by proper statistical diagnostics and are included only at the end of the iterative process. The method has already been successfully applied to segment man-made objects characterized by plane surfaces, like roofs, or by more complicated higher order geometry (Crosilla, Visintini and Prearo, 2004b). Nevertheless, it presents some critical aspects for the automatic extraction of the raw initial clusters, and for the extension of the process to the entire set of points, that contains also points not presenting any geometrical relationship with the particular cluster to be identified.
The paper proposes a new analytical method to automatically identify the initial raw data cluster relating to a generic geometrical feature. For every subset of homogeneous LIDaR data, the method identifies a limited number of surrounding points to submit to the refinement segmentation process, so to dramatically reduce computing time and memory. At the end, the algorithm makes it possible to automatically perform the segmentation of the entire data collection. The search of the initial homogeneous raw cluster of points is carried out by applying a local nonparametric regression algorithm (chapter 4), while the refinement process is performed by a robust parametric model, the before mentioned SAR one following the FS procedure.
For each LIDaR point, the nonparametric algorithm makes it possible to compute the predicted surface local trend value and its partial derivatives in the East and North directions. The LIDaR points belonging to the same homogeneous subset are characterized by a significant agreement between the measured and the predicted height and, for planar surfaces, by a further spatial constant value of the partial derivatives (chapter 5).

2. A SIMULTANEOUS AUTOREGRESSIVE SEGMENTATION MODEL

The proposed algorithm works under the hypothesis that LIDaR measures of the surface point height can be rightfully represented by the SAR model (Anselin, 1988):

$$\mathbf{z} - \rho\mathbf{W}\mathbf{z} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad (1)$$

where:

- $\mathbf{z}$ is the [n x 1] vector of laser height values (being n the total number of points to be segmented);
- ρ is a value (constant for the whole dataset) that measures the mean spatial interaction between neighbouring points;
- $\mathbf{W}$ is a [n x n] spatial adjacency (binary) matrix defined as $w_{ij} = 1$ if the points are neighbours, $w_{ij} = 0$ otherwise;
- $\mathbf{A}$ is a [n x r] matrix with $\mathbf{A}_i = \left[1 \quad E_i \quad N_i \quad \ldots \quad E_i^s \quad N_i^s\right]$ as rows where E_i and N_i are East and North-coordinates of points interpolated by a s = (r-1)/2 degree orthogonal polynomial;
- $\boldsymbol{\theta} = \left[\theta_0 \quad \theta_1 \quad \ldots \quad \theta_{r-1}\right]^T$ is a [r x 1] vector of parameters;
- $\boldsymbol{\varepsilon}$ is the [n x 1] vector of normally distributed errors (noise) with mean 0 and variance σ_ε^2.

To solve equation (1), a Maximum Likelihood (ML) estimation of the unknown parameters has been chosen. Let us start from

the following SAR log-likelihood function (Anselin, 1988), where **I** is the [n x n] identity matrix:

$$L(\boldsymbol{\theta},\rho,\sigma^2) = C + \ln|\mathbf{I}-\rho\mathbf{W}| - \left(\frac{n}{2}\right)\ln[(\mathbf{z}-\rho\mathbf{W}\mathbf{z}-\mathbf{A}\boldsymbol{\theta})^T(\mathbf{z}-\rho\mathbf{W}\mathbf{z}-\mathbf{A}\boldsymbol{\theta})] \quad (2)$$

The function (2) must be maximized not only with respect to $\boldsymbol{\theta}$ and σ^2, but also with respect to ρ. To avoid biased solutions this can be performed in stages (Pace, Barry and Sirmans, 1998), first by selecting a vector of length *f* of values over [0,1] labelled as $\rho_v = [\rho_1 \quad \rho_2 \quad \ldots \quad \rho_f]$ and then maximizing the profile log-likelihood function (for more details, see Crosilla, Visintini and Prearo, 2004a). The value ρ_{ML} giving the maximum log-likelihood value L is assumed as the ML estimation $\hat{\rho}$ of ρ. Finally, the optimal estimation of the SAR unknowns is obtained from:

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{I}-\hat{\rho}\mathbf{W})\mathbf{z} \quad (3.1)$$

$$\hat{\sigma}^2 = n^{-1}(\mathbf{z}-\hat{\rho}\mathbf{W}\mathbf{z}-\mathbf{A}\hat{\boldsymbol{\theta}})^T(\mathbf{z}-\hat{\rho}\mathbf{W}\mathbf{z}-\mathbf{A}\hat{\boldsymbol{\theta}}) \quad (3.2)$$

Regarding the order of the trend surface polynomial modelling $\mathbf{A}\boldsymbol{\theta}$, that is the r dimension of the $\boldsymbol{\theta}$ unknown parameter vector, the choice inferentially occurs by a *t*-Student test applied to the estimated $\hat{\boldsymbol{\theta}}$ values:

$$\frac{\hat{\theta}_i}{\hat{\sigma}_{\theta_i}} \le t_{(n-r)_{1-\alpha}}$$

where $\hat{\sigma}_{\theta_i}$ is the estimated standard deviation of $\hat{\theta}_i$, and α is the significance level of the test. In other words, once a redundant k-degree orthogonal polynomial (e.g. cubic, k = 3) has been assumed, the assessment of a reduced s < k degree, describing with plenty sensitivity the trend model, is then performed, so skipping not meaningful (k-s) parameters.

Within the **z** height values, the way to assess homogeneous spatial behaviour is to compute individual departures from the fitted polynomial trend surface. To this end, starting from (1), the vector $\mathbf{e} = \sigma^{-1}\boldsymbol{\varepsilon}$ of standardised residuals is computed as:

$$\mathbf{e} = \hat{\sigma}^{-1}[(\mathbf{I}-\hat{\rho}\mathbf{W})\mathbf{z} - \mathbf{A}\hat{\boldsymbol{\theta}}] \quad (4)$$

Afterwards, its n components are inferentially evaluated to find which measures do not fit the estimated trend surface: in fact, **e** is used to define the lack of fit statistic $\mathbf{e}^T\mathbf{e}$.
However, to robustly detect clusters of homogeneous stationary laser data, the estimations (3) and (4) have to be carried out by considering different subsets of the whole data set.

3. THE FORWARD SEARCH ALGORITHM AND THE STATISTICAL DIAGNOSTICS

An interesting algorithm to perform iterative SAR estimations on increasing datasets is the so-called "Block Forward Search" (BFS) proposed by Atkinson and Riani (2000). It makes possible to execute the robust estimations $\hat{\rho}$ and $\hat{\boldsymbol{\theta}}$ at each step of the search, starting from a partition of the dataset in blocks of contiguous spatial location, and considering them as elementary units of the entire set of points. In case of grid data, each block is a set of cells, while handling raw data it is difficult to univocally create the blocks: thus, the block dimension is merely unitary (UFS, Unitary Forward Search).
The basic idea of the FS approach is to repeatedly fit the postulated polynomial model to subsets of increasing size, selecting for any new iteration the observations **z** best fitting the previous subset, that is having the minimum standardised residual component in **e**. In equation (6), $\hat{\rho}$ and $\hat{\boldsymbol{\theta}}$ are estimated for each stationary cluster of data, while **z**, **A**, **W** and σ are referred to the whole dataset. Thanks to this growing strategy, the non-stationary data are included only at the end of the FS process, for each particular cluster.

In order to fix a rule to decide at which iteration the non-stationary data enter into the subset, a *F*-Fisher test is continuously carried out:

$$\frac{\left[\hat{\boldsymbol{\theta}}(m)-\hat{\boldsymbol{\theta}}(n_s)\right]^T\mathbf{A}^T\mathbf{A}\left[\hat{\boldsymbol{\theta}}(m)-\hat{\boldsymbol{\theta}}(n_s)\right]}{r\,\hat{\sigma}(n_s)^2} \le F_{(r,n-r)_{1-\alpha}}$$

For the generic *m*-dimensional subset of points, the null hypothesis states that the $\hat{\boldsymbol{\theta}}(m)$ values, estimated by (5.1), are not significantly different from the $\hat{\boldsymbol{\theta}}(n_s)$ values estimated with the n_s-dimensional initial subset. If such hypothesis is not satisfied, the just included point does not belong to the particular cluster since its presence provides a biased estimation of $\hat{\boldsymbol{\theta}}$.
Moreover, any new point included from now on can be classified as outlier or non-stationary data: from a strictly statistical point of view, there is no reason to continue with the iterations. It is then mandatory to define a mathematical rule or an operative routine to limit computations (3) and (4) only to the part of the entire dataset, where it is meaningful. Therefore, as it will be better explained in chapter 6, the algorithm has been designed so to repeat the segmentation process for all the geometrical features to be detected.

4. THE NONPARAMETRIC MODEL APPLIED FOR RAW SEGMENTATION

In order to suitably select raw homogeneous clusters from the whole dataset, we have exploited nonparametric regression techniques whose main quality is to allow the dependence analysis of a response variable on one or several predictors without specifying in advance the function relating the response to the predictors. In other words, this approach is a "data-driven" technique that determines the value of the regression function directly; therefore, a nonparametric analysis seems suitable for an exploratory use in the selection stage of a parametric model.
Three common methods of nonparametric regression are usually applied (Fox, 2004): *(nearest-neighbour) kernel estimation*, *local polynomial regression* and *smoothing splines*.
This paper proposes the use of a local polynomial regression that allows the weighted least squares estimation of a nonparametric smoother of the function and its two first order partial derivatives with respect to the East and North directions.
Given a response surface *z* over a domain $D \subset \Re^2$, consider the (k+1) differentiable mappings $\mu : D \to \Re$ and $\sigma : D \to \Re^+$. In addition let ε be a random variable with $E(\varepsilon) = 0$ and

$\mathrm{var}(\varepsilon) = 1$. The following model for an observation taken at the general location $\mathbf{x} \in D$, is assumed:

$$z(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon \quad (5)$$

where the partial derivatives of $\mu(\mathbf{x})$ exist and are continuous up to the order (k+1). Hence, to the extent of Taylor theorem, $\mu(\mathbf{x})$ can be approximated by a polynomial of order k in a neighbourhood of $\mathbf{x}$. Clearly, the functions $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ model a spatial deterministic component, while ε models a component of random noise. In this way, for each predicted point, the parameters of a local plane function can be determined. The analytical taylorized nonparametric model is (e.g. Sclocco and Di Marzio, 2004):

$$z_j = z_{0_i} + \left(\frac{\partial z}{\partial E}\right)_{E_i}(E_j - E_i) + \left(\frac{\partial z}{\partial N}\right)_{N_i}(N_j - N_i) + \varepsilon_j \quad (6)$$

where the weighted least squares estimates of z_{0_i}, i.e. the response surface value of z_i, and of the partial derivatives, i.e. the surface slope along East and North directions, are given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T\mathbf{Q}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Q}\mathbf{z} \quad (7)$$

where (for j = 1, …, p):

$$\boldsymbol{\beta} = \left[z_{0_i} \quad \left(\frac{\partial z}{\partial E}\right)_{E_i} \quad \left(\frac{\partial z}{\partial N}\right)_{N_i} \right]^T ; \quad \mathbf{A}_j = \left[1 \quad (E_j - E_i) \quad (N_j - N_i)\right]$$

and $\mathbf{Q}$ is a diagonal weight matrix defined by a symmetric unimodal kernel function centred on the i-*th* observation, whose diagonal terms are (Fox, 2004):

$$w_{ij} = \left[1 - \left(\frac{d_{ij}}{b}\right)^3\right]^3 \text{ for } \frac{d_{ij}}{b} < 1 \text{ and } w_{ij} = 0 \text{ for } \frac{d_{ij}}{b} \geq 1$$

where d_{ij} is the planimetric distance between j-*th* and i-*th* point, and b is the half-width of the window encompassing the p nearest neighbours of the i-*th* point.

According to Hardle (1990), the choice of the bandwidth, and not the choice of the kernel function, is critical for the performance of the nonparametric fit. The larger the value of b, the smoother the estimation of the regression function results, while the smaller the value of b, the larger the predicted point variance results. The need to balance bias and variance leads to the minimization of specific objective functions like, for instance, the *Prediction Sum of Squares* (Allen, 1974):

$$\mathrm{PreSS} = \sum_{j=1}^{n}(z_i - \hat{z}_{0i,-i})^2 = \min$$

where $\hat{z}_{0i,-i}$ is the fit of z at E_i, N_i when ignoring the observation z_i in obtaining the fit. This objective function tends to overfit the data by selecting a bandwidth too small, suggesting the possible need for penalizing functions of PreSS to protect against such small bandwidths (Hardle, 1990).

4.1 Local patterns of LIDaR data

From an exploratory point of view, local patterns of LIDaR data are quite different from the global spatial morphology exhibited by the whole region. In this regard, the statistical literature suggests the use of an adaptive bandwidth selection in such a way that the amount of smoothing changes according to the local complexity of the deterministic component. The smoother should also explicitly incorporate the structure of spatial dependence by including in its formulation the value of a contiguity criterion. This solution gives the smoothing a direction consistent with the a priori information on the spatial dependence, so avoiding an otherwise undifferentiated smoothing. In order to describe the local behaviour of a global spatial structure, Anselin (1995) suggested a class of indicators, the so-called "Local Indicators of Spatial Association" (LISA). Among these, we consider the Geary's local indicator:

$$c_i = \sum_{j \in J_i} \delta_{ij}(h_i - h_j)^2 \Big/ n^{-1}\sum_{i=1}^{n} h_i^2$$

where $h_i = z_i - n^{-1}\sum_{i=1}^{n} z_i$, while J_i denotes the set of the site labels for which the condition $\delta_{ij} > 0$, with $i \neq j$, holds.

Considering the [n x n] matrix $\boldsymbol{\Delta}$ defining a contiguity criterion, δ_{ij} is an entry of it, i.e. it indicates the spatially associated neighbour of the i-*th* location.

A typical criterion for constructing the contiguity matrix $\boldsymbol{\Delta}$ is the inverse distance or a common boundary. The above index is sensitive, first to a local cluster in a neighbourhood set of the i-*th* site (i.e. the set of the sites labelled by an element of J_i), second to spatial non-stationary data and outliers.

The Geary's indicator evaluates the spatial heterogeneity between the i-*th* height value and those belonging to a neighbourhood set. It can be used to compute an adaptive weight for the local bandwidth selection. Sclocco and Di Marzio (2004) propose the following weight term:

$$\gamma_i = c_i \sum_{j \in J_i} \delta_{ij} \Big/ \sum_{j \in J_i} \delta_{ij} c_j$$

that has to be multiplied by b, the original bandwidth value, to obtain an adaptive one. The combined effect is to reduce the amount of smoothing if the local morphology is complex.

5. THE IMPLEMENTED ALGORITHM

The C language program developed at the University of Udine for laser data processing (Beinat and Sepic, 2005) implements the previous local polynomial regression approach. The software directly handles the original ASCII raw data of irregular points, so that neither grid resampling nor a priori classification are required.

In particular, the following distinguished and independent steps characterize the algorithm:

1. Application of the nonparametric regression to the whole dataset of points;
2. Identification of homogeneous clusters of points as initial raw subsets to submit to SAR-FS parametric regressions;
3. Definition, for each cluster, of a further surrounding limited set of points: these sets constitute the searching areas for the SAR-FS parametric regressions.

Let us now analyse in detail each single step.

5.1 Nonparametric regression for the whole dataset

By applying the nonparametric regression, the 3-elements vector $\hat{\boldsymbol{\beta}}$ is computed for each point. The vector contains the

parameters of a local interpolating plane, estimated by means of the least squares regression process (7). To this purpose, a suitable bandwidth value is adopted. Its definition permits to limit the set of points involved in the computation of the above mentioned parameter vector. It is evident that a proper value of the bandwidth is fundamental for reaching reliable results, and that it is strictly joined to the LIDaR point density to which the segmentation process is applied. Anyway, once a starting value is fixed, the algorithm dynamically modifies its dimension so to automatically adapt it to the particular analysed surface.

5.2 Subset segmentation by a region growing method

To successfully apply the parametric regression, it is mandatory to identify zones for which the characteristic geometrical parameters have a homogeneous behaviour and where the most part of the data follows a specific trend. A suitable segmentation process of the whole dataset is then required. For such end, at this step, the algorithm creates various clusters starting from a randomly selected point not yet belonging to any other cluster. The surrounding points of the original chosen one are then analysed and the bandwidth value is exploited to limit the searching only to the points having a Euclidean distance less than the bandwidth. The clustering procedure is simultaneously carried out analysing the value of the predicted height $\hat{z}_{0_i}$ and of the partial derivatives along East $(\partial z/\partial E)_{E_i}$ and North $(\partial z/\partial N)_{N_i}$ directions, i.e. the slope of the interpolating planes. If the round points present difference values in slope and/or in height within a fixed threshold, than they are labelled as belonging to the same class and putted into a list. This algorithm goes on applying the same procedure to each list element, till this is fully completed.
Afterwards, the procedure runs again from the beginning, creating a new entity from a new point randomly chosen. The algorithm ends when every point has been analysed.
Summarising, by means of this method based on height difference values and slope evaluations, a first raw segmentation of the whole dataset is carried out: hence, each cluster of points will be a specific initial outlier-free subset for a SAR-FS parametric filtering process.

5.3 Definition of the searching areas

As mentioned before, each SAR-FS filtering process, starting from the just detected subsets, has to be confined to its surrounding points and not to the whole dataset. In other words, for each subset a searching area has to be identified. To this purpose, a Delaunay triangulation is accomplished for the whole dataset; afterwards, the points surrounding the subsets are analysed. If a point does not belong to any other subset and is closer than a fixed threshold, it is joined to the searching area. Once these operations are carried out for each subset, a corresponding number of searching areas are identified.

When the three fully automatic steps are ended, the whole dataset is subdivided into the same number of subsets and searching areas. In conclusion, being m the dimension of a generic feature within a dataset of $n >> m$ points, by means of this nonparametric algorithm, the achieved dimension of the raw subset n_s and of the searching area n_{sa} (with $n_s < m < n_{sa}$) is close enough to m. In this way, the iterations of the SAR-FS parametric algorithm are dramatically reduced (details about this software implemented by Matlab® in Crosilla, Visintini and Prearo, 2004a).

6. SOME APPLICATIONS

The mixed nonparametric and parametric segmentation algorithm has been tested on LIDaR data acquired with an Optech® ALTM 3033 airborne system over the City of Gorizia (North-East Italy) in November 2003 and April 2004. The helicopter scans of more than 30 million points are characterized by a mean density of 2 p.ts and 15 p.ts/m^2, respectively. Interested readers can find results and evaluations of the terrain parametric filtering method in Crosilla, Visintini and Prearo (2004a) and for building roofs detection in Crosilla, Visintini and Prearo (2004b). The latter problem of building modelling is widely described in the literature (e.g. see Brenner, 1999 for parametric algorithms, and e.g. Rottensteiner and Briese, 2003 for nonparametric ones). In the previous experiments, corresponding results have been obtained by processing the data with the SAR-FS program and the well-known TerraScan software (Soininen, 2003); this last software applies a nonparametric approach for the building modelling.

Throughout this chapter, some fully automatic applications of a mixed nonparametric and parametric segmentation method are presented: firstly, the case of a simple building is described in detail, while afterwards only uppermost results are shown for a very complicated edifice.
The first dataset consists of n = 29.662 high density acquired points over a building with a four roof planes near to a two roof planes house, a tree, a garden and a road ground surface, low vegetation and cars (see in Figure 1, the orthophoto (on left) and the axonometric view from South-West (on right)).

Figure 1. Left: orthophoto of one experimental area; right: view of laser points (colour by RGB from orthophoto).

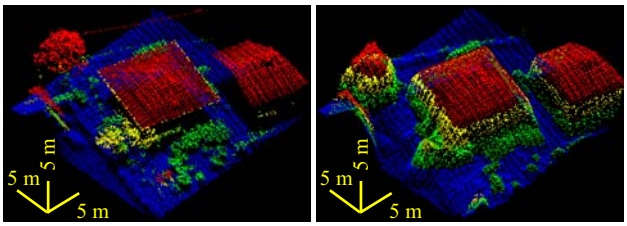

Figure 2. Left: view of laser points (colour by z_i elevation); right: view of interpolated points (colour by $\hat{z}_{0_i}$ elevation).

1) Nonparametric regression

In the first processing step, vector $\hat{\boldsymbol{\beta}}$ has been evaluated for each point. By comparing the estimated $\hat{z}_{0_i}$ values (Figure 2 on left), with the acquired z_i values (Figure 2 on right), the smoothing effect of the interpolation process can be clearly noticed. In particular in the upper part of Figure 2, let note how the aerial powerline, detected thanks to the high density scanning, yields a raising effect in the nonparametric interpolated surface. Figure 3 on left represents the values along the East direction of the planes slope: their range can vary from $-\infty$ to $+\infty$, especially around quasi-vertical elements, as the

building walls, the chimney pots and the tree boundary. Therefore, these extreme values are skipped in the plotting, leaving in black the corresponding areas.

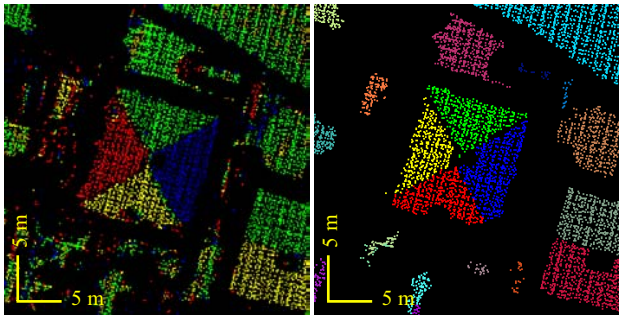

Figure 3. Left: plot of slope values along East (colour by value); right: raw subsets detected by region growing (colour by cluster).

2) Raw subset segmentation

In this step, 20 different clusters of points have been detected (Figure 3 on right). The only warning regards the mentioned points with extremely scattered slope: they have not been assigned to any cluster. In the same figure, let see how a lot of points are not yet assigned (black zones): apart from the quasi-vertical areas, these points belong to the irregular surface over the vegetation or to the roof ridges, namely everywhere the slope of the interpolated surface is very much variable.

3) Definition of the searching areas

In this step, for each one of the 20 previously selected subsets, the corresponding searching areas have been detected. In this way, the not yet segmented points fully contained in the raw subsets, as the chimneys, are assigned to the related searching areas, while the border area points (roof ridges) have a multiple allocation into each nearby area.

By means of some morphological considerations, the subsets associated to the roof planes can be identified with respect to those one relative to the ground. The four raw subsets and the corresponding searching areas of the case study building have been submitted to the parametric segmentation later described in detail. Figure 4 shows an axonometric view of the searching area points for the North roof plane, together with the vectors obtained by the specific tool "Construct Buildings" of TerraScan, for a better interpretation.

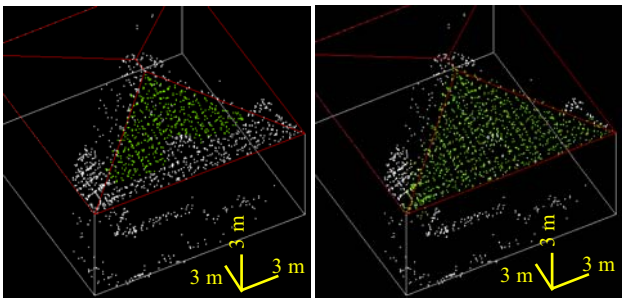

Figure 4. Points of searching area for North roof plane; left: raw subset (in green); right: SAR-FS final segmentation (in green).

4) Parametric segmentation

The 4 roof planes raw subsets (Figure 5 on left) have been submitted to the robust automatic SAR-FS segmentation (Figure 5 on right, and Figure 4 on right for the North roof plane only). The obtained results are very promising: the geometry of these plain surfaces is well determined, rightly rejecting each time the points belonging to chimneys, to contiguous planes or to the surrounding terrain.

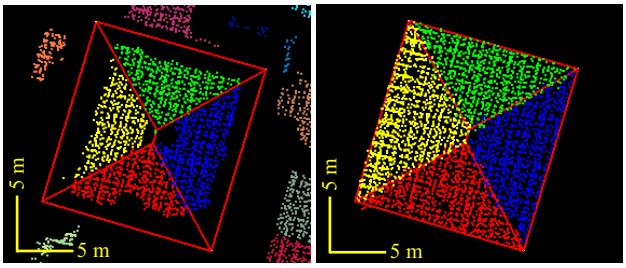

Figure 5. Left: raw subsets relating to roof planes; right: roof planes after parametric segmentation (both colour by cluster)

Figure 5 on right shows the final obtained segmentation of the roof points. It is in full agreement with the interactive assisted TerraScan modelling, so qualitatively proving the accuracy and reliability of the proposed mixed analytical procedure.

Furthermore to better understand the computational role of both nonparametric and parametric procedures, Table 6 reports the number of points involved in the segmentation process, starting from a dataset composed of 29.662 points.

Roof plane	points of raw subsets (n_s)	points of searching areas (n_{sa})	points found by SAR-FS (m)
North	644	1.993	1.220
East	749	1.777	1.511
South	664	2.022	1.069
West	517	2.398	1.601

Table 6: Number of points in the roof segmentation processes.

The nonparametric clustering averagely detects n_s = 643 points for each roof plane. The mean value n_{sa} of searching areas points is 2.047, while 707 is the mean number $(m - n_s)$ of points iteratively added to each plane by the SAR-FS procedure.

The second dataset consists of n = 75.288 points relating to an area with a large building, with 30 roof planes, a great tree in the court, and grass, road and pavement ground surface in the surroundings (see the orthophoto in Figure 7).

Figure 7: Orthophoto of a complex roof tested for segmentation.

The subsequent figures show the main aspects of the whole segmentation processing. The slope values along East direction present better regularity of the interpolated surface in correspondence of the roof and also for artificial ground areas (see Figure 8). Thus, by the region growing step, 41 different point clusters have been detected, as depicted in Figure 9 together with the TerraScan vector building modelling. Figure 10 and axonometric Figure 11 show the final nonparametric and parametric segmentation, performed for the 30 roof planes only. A good agreement of such a segmentation with the TerraScan modelling arises again, also for this rather complex building.

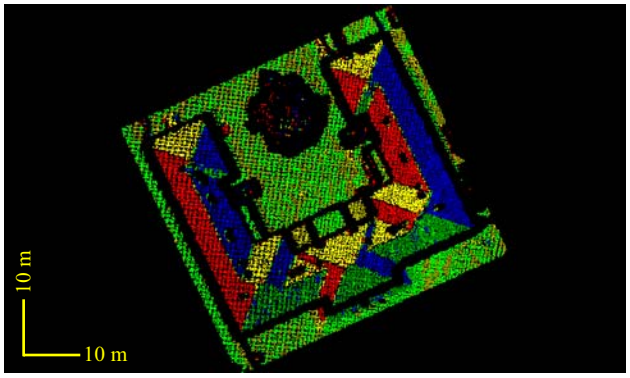

Figure 8. Plotting of slope values along East (colour by value).

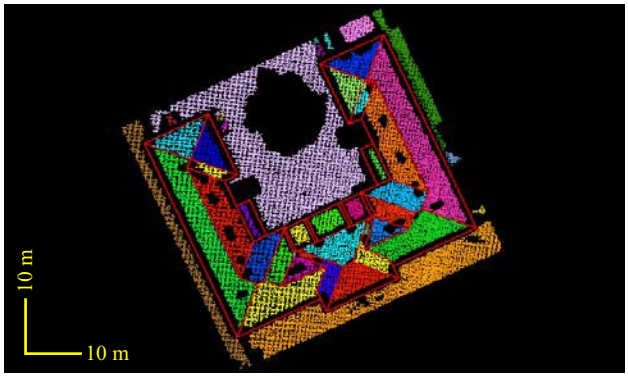

Figure 9. Subsets detected by region growing (colour by cluster).

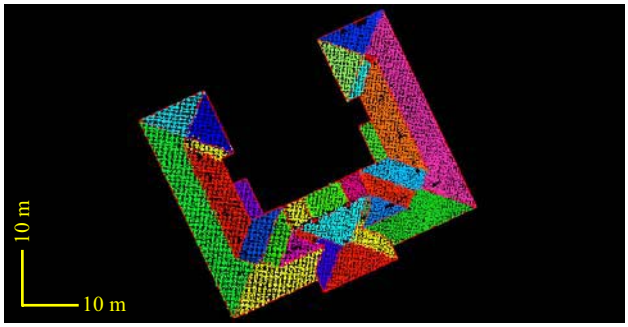

Figure 10. Final segmentation vs TerraScan modelling.

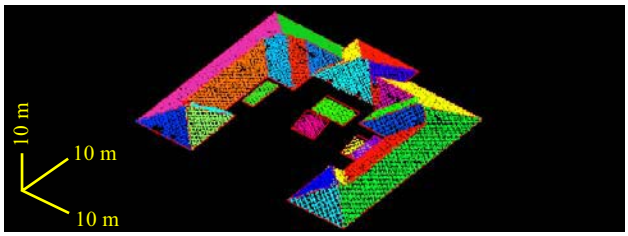

Figure 11. View of final segmentation vs TerraScan modelling.

7. CONCLUSIONS

The paper proposes an original technique for the robust segmentation of laser data. The method benefits of the combination of a mixed nonparametric and parametric regression model to automatically identify the homogeneous geometrical features present in the whole dataset. A local polynomial regression is used to define the original raw clusters of laser points. A forward search algorithm applied to a robust simultaneous autoregressive model is successively used to rigorously define the size and shape of the homogeneous geometrical features characterizing the laser data. The numerical results put in evidence the capability of the method proposed in terms of reduced computational memory, short time consuming and high definition and reliable segmented objects.

REFERENCES

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* 16, 125-127.

Anselin, L., 1988. *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht.

Anselin, L., 1995. Local indicators of spatial association - LISA, *Geographical Analysis*, 27, 93-115.

Atkinson, A.C., Riani, M., 2000. *Robust Diagnostic Regression Analysis*, Springer, New York.

Beinat, A., Sepic, F., 2005. Un programma per l'elaborazione di dati LIDaR in ambiente Linux (in Italian), *Bollettino della Società Italiana di Fotogrammetria e Topografia*, in press.

Brenner, C., 1999. Interactive modelling tools for 3D building reconstruction. *Photogrammetric Week 99*, Herbert Wichmann Verlag, Heidelberg, pp. 23-34.

Crosilla, F., Visintini, D., Prearo, G., 2004a. A robust method for filtering non-ground measurements from airborne LIDAR data, in: *Int. Arch.s of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXV, B3, Istanbul, 196-201.

Crosilla, F., Visintini, D., Prearo, G., 2004b. Sperimentazione in ambito urbano dell'algoritmo autoregressivo SFS (Spatial Forward Search) per il filtraggio dei dati laser (in Italian), *Atti dell'VIII Conferenza Nazionale ASITA*, 1, 937-942.

Fox, J., 2004. *Nonparametric Simple Regression: Smoothing Scatterplots*. Sage, Thousand Oaks.

Hardle, W., 1990. *Applied Nonparametric Regression*, Cambridge University Press.

Pace, R.K., Barry, R., Sirmans, C.F., 1998. Quick computation of spatial autoregressive estimators, *Journal of Real Estate Finance and Economics*, 1-12.

Rottensteiner, F., Briese, C., 2003. Automatic generation of building models from LIDAR data and the integration of aerial images. in: *Int. Arch. of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Dresden, Germany, Vol. XXXIV, Part 3/W13, pp. 174-180.

Sclocco, T., Di Marzio, M., 2004. A weighted polynomial regression method for local fitting of spatial data, *Statistical Methods & Applications*, 13, 315-325.

Sithole, G., Vosselman, G., 2003. Comparison of filtering algorithms, in: *Int. Arch. of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIV, B/W13, Dresden, 71-78.

Soininen, A., 2003. *TerraScan*. User Guide. Terrasolid.

ACKNOWLEDGEMENTS

This work was carried out within the research activities supported by the INTERREG IIIA Italy-Slovenia 2003-2006 project "Cadastral map updating and regional technical map integration for the GIS of the regional agencies by testing advanced and innovative survey techniques".