Quality assessment for automatic LiDAR data classification methods

André Caceres Carrilho[1]
Ivana Ivánová[2]
Mauricio Galo[2]

Universidade Estadual Paulista – UNESP
[1] Programa de Pós-Graduação em Ciências Cartográficas - PPGCC
[2] Departamento de Cartografia
Rua Roberto Simonsen, 305 – 19060-900 – Presidente Prudente - SP, Brasil
carrilho.acc@gmail.com, {i.ivanova, galo}@fct.unesp.br

Abstract. This paper provides an initial discussion on standardization of quality assessment of thematic accuracy of classification methods applied to LiDAR (Light Detection And Ranging) data. The literature review exposes an overall lack of consensus for quality control regarding LiDAR point clouds and derived products. To mitigate this problem, the information retrieval theory is reviewed and a case study is presented aiming at the thematic accuracy analysis that precision, recall and F-score elements can provide. Fitness for use is discussed focusing on the selection of spatial data quality elements for practical applications, and an approach for algorithm evaluation is presented. Although many alternatives can be considered in solving this problem, some directions are appointed in order to continue the research.

Keywords: Quality control, LiDAR processing, automatic classification .

1. Introduction

There are topics in the remote sensing and photogrammetric literature related to spatial data quality and LiDAR data processing with great interest in both theoretical and practical aspects. The most recurrent is the positional accuracy of the point coordinates, which is a study object during data collection process, using covariance propagation models from the raw data to the final point cloud. During data processing, most of the tasks (such as filtering, classification, and reconstruction) are often performed with automatic methods. The usage of those methods is justified by the nature of the point cloud data (files with up to billions of laser returns), which makes it impractical to execute data processing manually.

A statistical evaluation of an agreement between the obtained results and the reality is necessary for validation of any automatic method of spatial data process. In the literature there are several proposed methods regarding automatic filtering of ground points (laser returns belonging to the ground surface), classification and extraction of features from LiDAR data. Most of these methods differ in the means used to report the quality of the results, in which the lack of consensus for quality control is evident. Part of the disagreement is related to the criteria used during the selection of the areas to perform the analysis and which metrics to consider for this evaluation.

The main objective of this paper is to discuss and provide a suitable approach to perform quantitative (statistical) evaluation in LiDAR data classification results. Additionally, focusing on practical applications, we propose a method to evaluate classification algorithms. The results of this study appoint to possible future developments of automatic quality control and internal validation of thematic accuracy and overall quality of classified LiDAR data.

1.1. Absence of a consensus for quality control in processing of LiDAR data

Vieira and Mather (2005) presented a review of standard methods to assess the quality (for both positional and thematic accuracy) of cartographic products obtained from remotely sensed data. Correspondence between features in the imagery and the reference data was the measure for positional accuracy, while the thematic accuracy assessment was based on statistics derived from the confusion (error) matrix.

Specifically in the airborne laser scanning (ALS) scope, Sithole and Vosselman (2004) evaluated the performance of eight filter algorithms for bare-Earth extraction from point clouds. A set of point clouds was filtered with all algorithms and the results were compared to the reference data, generated by filtering the entire datasets using aerial photographs for inspection. The type I (rejection of ground returns) and type II errors (acceptance of non-ground returns) were used in the quantitative comparison of the filters.

Li, Xiao and Wang (2013) introduces a method to assess the quality of building roof plane segmentation generated from LiDAR data using the random sample consensus (RANSAC) algorithm. The strategy proposed by the authors differs from previous approaches by only using in the analysis a reduced number of samples instead of the entire dataset. The selection criteria for the buildings used in the evaluation of the method was based in the geometric complexity (shape) of the roof.

In Buján et al. (2012), however, the quality was assessed by randomly selecting portions of the classified point clouds and evaluating the level of agreement of the results with the reference (manually identified buildings) in the data. The classification approach, in this case specific for rural area, uses a hierarchical object-based method, combining LiDAR data and aerial images.

For ground filtering algorithms, Meng, Currit and Zhao (2010) discussed the challenge on quantitative accuracy assessment. The authors provide an analysis of several published methods that differs both in the procedure used to generate ground truth data and the spatial data quality elements. About half of the papers listed in Meng, Currit and Zhao (2010) review only performed visual inspection, and most of which compute quantitative elements based on type I and type II errors to evaluate the digital terrain model (DTM).

The lack of standard methodologies for the quality control of LiDAR point clouds is reported in Habib el al. (2010), which provides internal quality control procedures to evaluate positional accuracy of LiDAR data. The method is based on the iterative closest point (ICP) method and assumes that the data was captured in parallel flight lines with overlap.

2. Classification results and thematical accuracy

As shown in Vieira and Mather (2005), a general approach to assess the thematic accuracy of classified data is the computation of a confusion matrix, which compares the results of a classification method with the reference data, and then measuring the closeness of the attributes to the truth using various elements (comission and omission errors, for instance).

Performing a per-point analysis of the classification results with respect to a class (k) four cases can occur according to Buján et al. (2012):

- True Positive (TP) if the considered point and its correspondent in reference data belong to the same class k.
- False Positive (FP) when the computed point label was k although the corresponding in the reference is different.
- True Negative (TN) if both resulting and reference classes are different from k.
- False Negative (FN) when the resulting point class is different from the reference, which is equals to k.

In statistical inference, the false positive and false negative values are equivalent to type I and type II errors, respectively. This nomenclature derives from hypothesis tests theory: the type I error is defined as the rejection of the null hypothesis when it is true, and the type II error is the failure to reject the null hypothesis when it is false.

The precision and recall elements, known from the fields of pattern recognition and information retrieval are often used for evaluation of classification and regression algorithms. Although these two elements can be computed directly from the data, the precision and recall values are usually derived from the confusion matrix. If the true positive, false positive, and false negative rates are known to a specific class, the precision (p) and recall (r) values for this class can be computed as:

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN} \tag{1}$$

Although the nomenclature for these two elements is known and widely used in the information retrieval research field, both photogrammetry and remote sensing communities adopted the terms correctness and completeness as synonyms for precision and recall, respectively. Besides these indicators, in other knowledge areas other parameters and terms can be used to quantify the quality or performance of a classification. As an example we can mention the sensitivity (equivalent to recall) and specificity, as can be seen in Sokolova, Japkowicz and Szpakowicz (2006). The remote sensing literature also presents other quantities derived from the confusion matrix, such as the branching and miss factors, and the detection and overall quality percentages of the classification (see Hermosilla et al. (2011), Buján et al. (2012), and Awrangjeb and Fraser (2014)).

To simplify the analysis, some authors (Sokolova, Japkowicz and Szpakowicz (2006), Lu et al. (2014), and Vega et al. (2014), for instance) adopt a composite value called F-score, computed as the weighted harmonic mean of precision (p) and recall (r) elements. The F-score is a useful measure because it takes into account the three quantities TP, FP and FN, used in the precision and recall computation. When the precision and recall weights are equal, the F-score computation can be simplified to:

$$\text{F-score} = \frac{2}{1/p + 1/r} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

2.1. Data quality elements and fitness for use

The Technical Committee 211 of the International Organization for Standardization defines several spatial data quality elements in the ISO 19157 (2013) standard, including completeness, logical consistency, positional accuracy, thematic accuracy, temporal quality, and usability. Selecting suitable spatial data quality elements for the analysis is the first step of the quality evaluation process, and this selection should be based on the purpose and usability of the product, for a given application.

As examples of selection of suitable spatial data quality elements, let us consider the cadastral survey and property register of real state, where the positional accuracy of the property boundaries has a important role. For navigation tasks, besides the positional accuracy, the orientation accuracy is other important element to be considered. In navigation application the tolerance in positional accuracy should be more flexible when comparing with the first application, since current real-time positioning devices only provide users locations within a margin of error of 5 m to 10 m.

Besides the standard spatial data quality elements other characteristics related to the acquisition can be evaluated, for instance, in cases where area calculations are important, the

point cloud density is a key element that measures the level of detail that both surface and edges of the building were sampled. This value can be used to predict the uncertainty of area calculations from the data (by applying covariance propagation models), and it also may provide information on how discernible two ground targets are.

The precision, recall, and F-score values are reasonable data quality measures to evaluate thematic accuracy, since they demonstrate the compliance of the results with the reference data. Those elements give an overall perspective of the classification, when a particular and detailed analysis might not be feasible. For instance, if the evaluated method tends to misclassify building roof borders as vegetation, that can be caused by the interpolation applied during the grid generation step (see Figure 1), and it would not be possible to conclude this using only precision and recall values.

2.2. Evaluating algorithms

For practical applications the quality of classification algorithms should be evaluated. In a ideal scenario, the algorithms must be in conformance with standards. For instance, American Society for Photogrammetry and Remote Sensing (ASPRS) defines the LAS file format for interchange of 3-dimensional point cloud data, and it is expected that any service using and/or producing LiDAR data supports this format. Heidemann (2014) presents a base scheme for LiDAR point cloud classification, the format requirements and also the maximum percentage of errors within a specified area.

Besides achieving reliable results, other two major concerns in any application are the performance and scalability of classification algorithms, i.e. the rate that the computational resources require to solve the problem and the relation between processing time with respect to dataset's size. The performance and scalability can be measured in terms of algorithm complexity, relating the number of operations with the dataset's size.

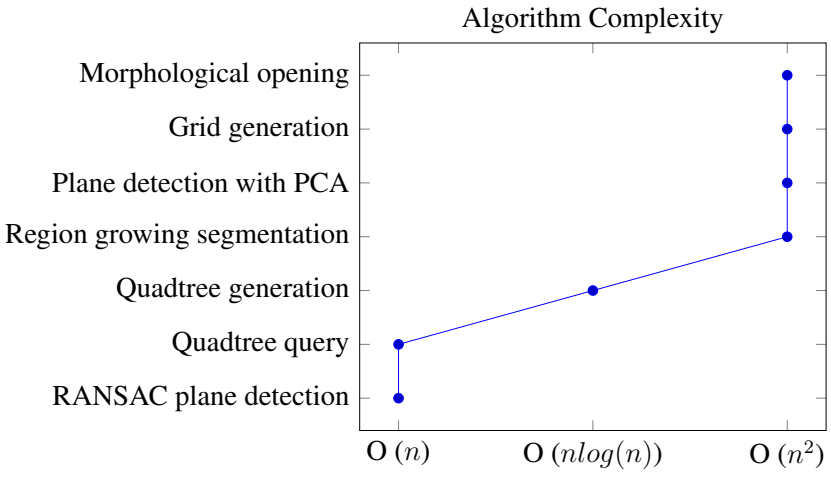


Figure 1: Approximate complexity of the algorithms used by the classification method described in Carrilho (2016).

As can be seen in Cormen et al. (2001), a common notation used in the computer science to assess algorithm complexity is the "Big O" or O-notation. This notation is independent from the programming language, implementation details, and hardware used, thus facilitating the selection of a suitable algorithm for a given problem. Most algorithms presented in Figure 1 tend to $O(n^2)$ complexity, which indicates that they will average to a quadratic increase in processing time and computational resources. In this sense, a quadratic raise in processing time

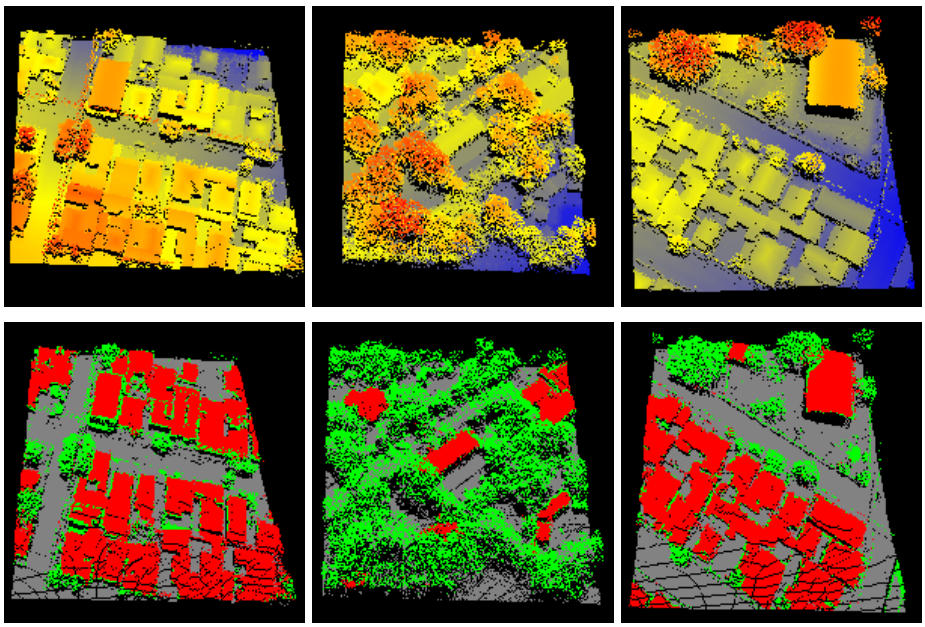
and memory usage over the increment in dataset size is expected. A more detailed description of the algorithms can be seen in Carrilho (2016).

Aside from the complexity analysis, it is necessary to take into account other factors. For instance, the grid generation for digital surface models (DSM) is preferred for tasks where the processing time is crucial (real-time applications, for instance) since the raster format allows direct access to point heights given they are regularly spaced. However, since the grid generation requires interpolation of the original heights, the process increases uncertainty in the data, thus, the impact of using a raster structure in the classification method must be studied. The grid structure is not suitable for vegetation studies and dendrometric computations, since they usually require all LiDAR returns.

3. Thematic accuracy of Presidente Prudente LiDAR data classification

This case study aims at evaluation of LiDAR data classification results obtained by the method used in our previous work (CARRILHO, 2016), that uses mathematical morphology (MM) concepts to distinguish which are the ground returns, defining the DTM, principal component analysis (PCA) for a rough approximation of planar regions, and the random sample consensus (RANSAC) algorithm for extracting building roof planes following the sequence shown in Figure 1. The dataset was acquired by Sensormap Geotenologias in December 2014 from a flight over Presidente Prudente city, using RIEGL® LMS Q680i airborne laser scanning. The raw files were made available in LAS file format version 1.2, and presents an average density of 14,9 pulses/m^2.

To assess the thematic accuracy of the classification method, three square regions of 10,000 m^2 were selected. Differently from the criteria used by Li, Xiao and Wang (2013), the selection of the regions (Figure 2) was made considering general characteristics of the features.



(a) Region 1 (b) Region 2 (c) Region 3

Figure 2: Perspective projection view of the selected regions (top) and respective classification results (bottom).

The first region (a) covers a residential area with small buildings and a few trees, while the second region (b) contains mostly vegetation. The third region (c) presents small houses and a

large hangar with some trees nearby. In Figure 2 the top images correspond to the original point cloud and the bottom images show the respective regions after classification.

A reference dataset was generated manually for each region using high resolution aerial images for inspection. The classification algorithm was applied with the same configuration to the three regions and the comparison of the output with the reference data enabled the computation of precision and recall values for each class, as presented in Figure 3.

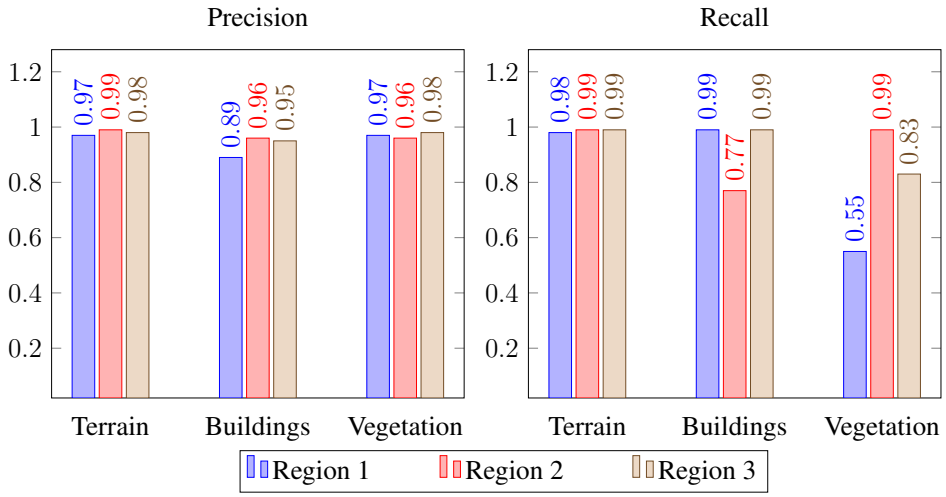


Figure 3: Precision and recall values computed for the three regions.

The recall values of regions 1 and 3 for the vegetation class, indicates lack of robustness of the method for tree detection. The same occurs with the recall value of region 2 for the buildings class, where some tree canopies were wrongly labeled as building roofs. This confusion is assigned to the thresholds used in the PCA pre-classification step which considers homogeneous portions of the canopies as planes and the RANSAC algorithm fails to remove them.

For this case study we also computed F-score, and as it can be seen in Figure 4, it has similar values as the recall. However, the F-scores present less variation than the recall values as a result of precision values being considered in the computation of the harmonic mean (Eq. 2).

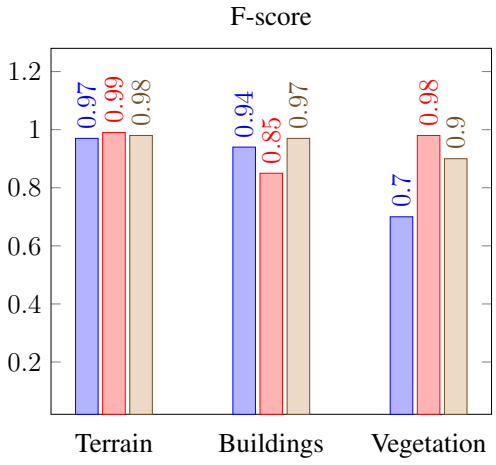


Figure 4: F-score values computed for the three regions.

4. Conclusions

In this paper, we presented the current state in LiDAR post processing quality control, which show an absence of consensus for automatic classification methods. We also discussed ways to evaluate the complexity of classification algorithms using mathematical measures. To provide brief analysis on data quality elements, a case study was performed using an automatic classification method and in one portion of a point cloud from Presidente Prudente city. The precision, recall and F-score values were computed for the classes using manually generated datasets as ground truth. The results indicate that these elements are reasonable choices to assess the overall level of agreement of classification with the reference data. However, an in-depth analysis of the results must consider other parameters to achieve detailed conclusions.

The data quality elements discussed in this paper can be used for internal validation, however, a problem that remains unsolved to perform automatic quality control is the acquisition of reference data compatible to the classification results. There are alternatives to obtain reference data, for instance results from commercial softwares or different classification methods. This solution, however, is not always affordable, therefore limiting the possibilities of continuous development. The future research related to this study comprises alternative approaches to perform automatic quality control of classification results.

Acknowledgement

The authors would like to acknowledge the support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the scholarship, from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the support to the project n. 307788/2012-1, and Sensormap Geotecnologias, which provided the data used in this work.

References

AWRANGJEB, M.; FRASER, C. S. Automatic segmentation of raw lidar data for extraction of building roofs. *Remote Sensing*, v. 6, n. 5, p. 3716, 2014. ISSN 2072-4292. Disponível em: <http://dx.doi.org/10.3390/rs6053716>.

BUJÁN, S.; GONZÉLES-FERREIRO, E.; REYES-BUENO, F.; BARREIRO-FERNANDÉZ, L; CRECENTE, R.; MIRANDA, D. Land use classification from lidar data and ortho-images in a rural area. *The Photogrammetric Record*, Blackwell Publishing Ltd, v. 27, n. 140, p. 401–422, 2012. ISSN 1477-9730. Disponível em: <http://dx.doi.org/10.1111/j.1477-9730.2012.00698.x>.

CARRILHO, A. C. *Application of image processing and analysis techniques towards buildings and vegetation detection using lidar data*. Tese (dissertation) —São Paulo State University (UNESP), 2016. Disponível em: <http://repositorio.unesp.br/handle/11449/137752>.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. *Introduction to Algorithms*. 2nd. ed. [S.l.]: McGraw-Hill Higher Education, 2001. ISBN 0070131511.

HABIB, A. KERSTING, A. P.; BANG, K. I. Alternative methodologies for the internal quality control of parallel lidar strips. *IEEE Transactions on Geoscience and Remote Sensing*, v. 48, n. 1, p. 221–236, Jan 2010. ISSN 0196-2892. Disponível em: <http://dx.doi.org/10.1109/TGRS.2009.2026424>.

HEIDEMANN, H. K. Lidar base specification (ver. 1.2, november 2014). In: JEWELL, S.; KIMBALL, S. M. (Ed.). *U.S. Geological Survey Techniques and Methods, book 11, Collection and Delineation of Spatial Data, chap. B4*. Reston, Virginia: United States Geological Survey (USGS), 2014. p. 67. Disponível em: <https://dx.doi.org/10.3133/tm11B4>.

HERMOSILLA, T. RUIZ, L. A.; RECIO, J. A.; ESTORNELL, J. Evaluation of automatic building

detection approaches combining high resolution images and lidar data. *Remote Sensing*, v. 3, n. 6, p. 1188, 2011. ISSN 2072-4292. Disponível em: <http://dx.doi.org/10.3390/rs3061188>.

ISO 19157. *Geographic information — Data quality*. International Organization for Standardization, Geneva, CH, dec 2013.

LI, J.; XIAO, Y.; WANG, C. Quality assessment of building roof segmentation from airborne lidar data. In: *2013 21st International Conference on Geoinformatics*. [s.n.], 2013. p. 1–4. ISSN 2161-024X. Disponível em: http://dx.doi.org/10.1109/Geoinformatics.2013.6626195>.

LU, X.; GUO, Q.; LI, W.; FLANAGAN, J. A bottom-up approach to segment individual deciduous trees using leaf-off lidar point cloud data. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 94, p. 1–12, 2014. ISSN 0924-2716. Disponível em: <http://dx.doi.org/10.1016/j.isprsjprs.2014.03.014>.

MENG, X.; CURRIT, N.; ZHAO, K. Ground filtering algorithms for airborne lidar data: A review of critical issues. *Remote Sensing*, v. 2, n. 3, p. 833, 2010. ISSN 2072-4292. Disponível em: <http://dx.doi.org/10.3390/rs2030833>.

SITHOLE, G.; VOSSELMAN, G. Experimental comparison of filter algorithms for bare-earth extraction from airborne laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 59, n. 1–2, p. 85 – 101, 2004. ISSN 0924-2716. Advanced Techniques for Analysis of Geo-spatial Data. Disponível em: <http://dx.doi.org/10.1016/j.isprsjprs.2004.05.004>.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: *Proceedings of the AAAI'06 workshop on Evaluation Methods for Machine Learning*. [S.l.: s.n.], 2006. p. 24–29.

VEGA, C.; HAMROUNI, A.; MOKHTARI, S. E.; MOREL, J.; BOCK, J. RENAUD, J. P.; BOUVIER, M.; DURRIEU, S. Ptrees: A point-based approach to forest tree extraction from lidar data. *International Journal of Applied Earth Observation and Geoinformation* v. 33, p. 98–108, 2014. ISSN 0303-2434. Disponível em: <http://dx.doi.org/10.1016/j.jag.2014.05.001>.

VIEIRA, C. A. O.; MATHER, P. M. Techniques for estimating the positional and thematic accuracy of remotely sensed products. In: *Anais do XII Simpósio Brasileiro de Sensoriamento Remoto.* v. 1, 2005. p. 4351–4359. Disponível em: <http://marte.sid.inpe.br/col/ltid.inpe.br/sbsr/2004/11.03.15.47/doc/4351.pdf>.