Expert Systems With Applications 230 (2023) 120672

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# A discordance analysis in manual labelling of urban mobile laser scanning data used for deep learning based semantic segmentation

Silvia María González-Collazo [a,*], Jesús Balado [a], Elena González [a], Abdul Nurunnabi [b]

[a] CINTECX, Universidade de Vigo, GeoTECH Group, 36310 Vigo, Spain
[b] Department of Geodesy and Geospatial Engineering, FSTM, University of Luxembourg, L-1359, Luxembourg

ARTICLE INFO

*Keywords:*
Deep learning
LiDAR
Machine learning
Mobile mapping systems
Point clouds
Semantic segmentation

ABSTRACT

Labelled point clouds are crucial to train supervised Deep Learning (DL) methods used for semantic segmentation. The objective of this research is to quantify discordances between the labels made by different people in order to assess whether such discordances can influence the success rates of a DL based semantic segmentation algorithm. An urban point cloud of 30 m road length in Santiago de Compostela (Spain) was labelled two times by ten persons. Discordances and its significance in manual labelling between individuals and rounds were calculated. In addition, a ratio test to signify discordance and concordance was proposed. Results show that most of the points were labelled accordingly with the same class by all the people. However, there were many points that were labelled with two or more classes. Class *curb* presented 5.9% of discordant points and 3.2 discordances for each point with concordance by all people. In addition, the percentage of significative labelling differences of the class *curb* was 86.7% comparing all the people in the same round and 100% comparing rounds of each person. Analysing the semantic segmentation results with a DL based algorithm, PointNet++, the percentage of concordance points are related with *F-score* value in $R^2 = 0.765$, posing that manual labelling has significant impact on results of DL-based semantic segmentation methods.

## 1. Introduction

Deep Learning (DL) approaches have recognized as the state-of-the-art-in 3D point cloud classification and semantic segmentation (Zhang et al., 2019). The quality and accuracy of results of the supervised DL approaches are highly dependent on the training data. Although DL approaches do not yet achieve perfect segmentation results for complete scenarios, they have already achieved very high success rates on many tasks. For example, building floors were segmented with 98.5%, 98.2%, and 98.1% of IoU (Intersection over Union) using PointWeb (Zhao et al., 2019), LLGF-Net (Zhang et al., 2022), and HybridCR (M. Li et al., 2022) methods, respectively. Roads were segmented with 98.5% of IoU using CGA-Net (Lu et al., 2021), vegetation were segmented with 97.3% of IoU using MinkowskiNet (Choy et al., 2019), and buildings were segmented with 97.5% of $F_1$-Score using ResNet18 algorithms.

Usually, and for a reasonable comparison, methods are tested on public datasets. There are numerous point cloud datasets covering a wide variety of scenarios, outdoor including urban (Ros et al., 2016) and road (Behley et al., 2019), and indoors (Armeni et al., 2017). Most of this data is acquired with mobile mapping systems on account of its good relationship between acquisition time and accuracy (Ma et al., 2018). These datasets are composed of two main pieces of information. They can be categorized as labelled and unlabelled data. The raw data representing the 3D environment, usually unlabelled, which can be acquired by laser scanners or are generated synthetically (Deschaud et al., 2021). For the labelled data, the labels associated to each point provide corresponding object class that can serve as the ground truth for training, validation and testing (Xie et al., 2020) purposes in supervised machine learning and deep learning methods. The data labelling is mainly done manually, so labelling is a subjective task and may contain human errors. Visualization of point clouds is difficult, mainly because of irregular point density, (Richter & Döllner, 2014), seeing through the surface effect (Virtanen et al., 2020), complex 3D shapes (Uchida et al., 2020) and sometimes, absence of true colour (González et al., 2022).

The objective of this study is to quantify discordances between the labels made by different people in order to assess whether such discordances can influence the success rates of a DL based semantic segmentation algorithm. The study used an urban point cloud covering 30 m of

* Corresponding author.
*E-mail addresses:* silvgonzalez@uvigo.es (S.M. González-Collazo), jbalado@uvigo.es (J. Balado), elena@vigo.es (E. González), pyal1471@yahoo.com (A. Nurunnabi).

https://doi.org/10.1016/j.eswa.2023.120672
Received 14 December 2022; Received in revised form 20 April 2023; Accepted 30 May 2023
Available online 2 June 2023

area, since urban areas present high complexity and overlapping of objects. Besides urban areas are one of the main scenarios of application of LiDAR technology. From a set of instructions, ten researchers familiarized with point clouds performed their corresponding labelling with two repetitions. Based on the results and considering the discordances, a ground truth can be obtained based on a voting system.

The reminder of this paper is structured as follows. In Section 2, semantic segmentation works are briefly reviewed. The proposed method is explained in Section 3. Section 4 presents and analyse the results. The discussion is in Section 5, and Section 6 concludes this paper.

## 2. Related work

Deep Learning methods are widely used in several areas, such as health care, visual recognition, self-driving cars, classification of the terrain. Abdelaziz Ismael et al., (2020) presented a method for classifying brain tumour types based on Magnetic Resonance Images. Hyperspectral images are also used in deep learning methods, (Hong et al., 2022). They proposed a novel backbone network called SpectralFormer, to extract and represent the attributes of spectral signatures with more accuracy. In (S. Li et al., 2019) the authors presented a review of deep learning based Hyperspectral image classification, comparing several strategies of this topic. However, most of the deep learning techniques are designed and applied for a single modality. Hong et al., (2021) proposed a method to identify the materials lying over or beneath the Earth's surface, developing a multimodal deep learning (MDL) framework. Deep learning methods are also used to classify point clouds. Labelled point clouds are widely applied to train and test DL methods, both in indoor and outdoor environments. Some researchers process point clouds through a rasterization (Guiotte et al., 2020; Lê et al., 2022; Paz Mouriño et al., 2021) and voxelization (Tchapmi et al., 2017; Xu et al., 2021). In this way, advantages from well-studied Convolutional Neural Networks (CNNs) can be applied to point cloud structures. Today, most DL based studies for semantic segmentation in point clouds are based on point Neural Networks, such as PointNet (Charles et al., 2017), PointNet++ (Qi et al., 2017a) or their variants (Boulch, 2020; Hu et al., 2019; Nurunnabi, Teferle, Li, Lindenbergh, & Hunegnaw, 2021; Thomas et al., 2019).

Point cloud semantic segmentation is frequently applied for intelligent city modelling, autonomous driving and urban planning. Nurunnabi, Teferle, Li, Lindenbergh, & Parvaz (2021) investigated PointNet algorithm and its potential for large-scale outdoor point clouds. The authors in Balado et al. (2019) applied PointNet for semantic segmentation of road environments in 7 classes. In Liu et al. (2022), the authors proposed a novel context-aware network (CAN) that can directly deal with large-scale point clouds. They tested this architecture in three datasets, Tongji-3D dataset (Chun et al., 2021), Semantic 3D dataset (Hackel et al., 2017) and S3DIS dataset (Armeni et al., 2016), obtaining a mean Intersection over Union (IoU) value of 83.52%, 74.7% and 78,97% respectively. A training dataset for sematic segmentation of urban point cloud map for intelligent vehicles is presented in Song et al. (2022). The urban area was scanned with a mobile mapping system. Then, they performed a semi-automatic and manual labelling. To verify the training dataset, various state-of-the-art point cloud semantic segmentation algorithms were used, obtaining a mean IoU value of 53.7% (ConvPoint (Boulch, 2020)), 52.8% (KPConv (Thomas et al., 2019)), 45.8% (SPG (Landrieu & Simonovsky, 2018)), 46.58% (RandLA-Net (Hu et al., 2019)) and 25.5% (PointNet++ (Qi et al., 2017a)).

Some authors provide large-scale urban Mobile Laser Scanning (MLS) datasets, with large number of labelled points. iQumulus dataset (Vallet et al., 2015) is one of them composed by 10 km of streets in Paris. This data set consists of 300 million points labelled with 11 classes. Semantic3D dataset (Hackel et al., 2017) is composed by 4.5 km area, having 4,000 million of points with 8 class labels (man-made terrain: mostly pavement, natural terrain: mostly grass, high vegetation: trees and large bushes, low vegetation: flowers or bushes smaller than 2 m, buildings, remaining hard scape, scanning artifacts and cars and trucks). In Roynard et al. (2018), authors presented an urban dataset composed by several point clouds of outdoor scenes in Paris and Lille of 1.94 km length, and 143 million hand labelled points of more than 50 classes (e.g., the ground, cars and benches). A large-scale urban outdoor point cloud dataset acquired by an MLS system in Toronto is showed in Tan et al. (2020). This dataset has 1 km length, 78 million points and 8 classes. In Zhu et al. (2020), the authors presented a dataset which covers an urban area in Munich of approximately 1 km long and include more than 40 million labelled points with 8 classes. Generation of datasets has reached also case studies with limited data samples. ArCH (Arquitectural Cultural Heritage) (Pierdicca et al., 2020) is composed by 11 labelled point clouds involving both indoor and outdoor scenes have points of churches, chapels, cloisters, porticoes and arcades.

The previous works compare the semantic segmented result with the ground truth (i.e. manual labels). However, none of the studies measures the discordance in the labelled ground truth, which can influence the results of DL methods. The novelty of this research is the measurement of discordance in manual labelling in MLS point clouds and the discordance influence in DL based semantic segmentation methods.

## 3. Materials and methods

### 3.1. Input data

An MLS point cloud used in this study was acquired by Riegl VUX-1HA, which is a compact and lightweight laser scanner. The technical specifications of the equipment are given in Table 1. The three-dimensional (3D) point cloud $P(X, Y, Z)$ of an urban road contains 4 million points, with a density of 6,000 points/m$^2$ without any attribute. It covers 30 m of a street in the city of Santiago de Compostela (Spain). The quantity of data was considered representative for an urban scene due to the same pattern is repeated along the roads. The point cloud is a part of the Santiago Urban Dataset (SUD) dataset (segment H) (González-Collazo et al., 2022). This area was selected because it contains highly representative characteristics of an urban geometry. It composed of a two-line road, one side parking area with few cars, sidewalk on both sides, building facades with entrances, trees, streetlights, traffic signs and street furniture.

### 3.2. Manual labelling

The point cloud was labelled manually by 10 persons ($i = 1, 2, \cdots, 10$) and in two rounds ($r = 1$and$2$), generating each one the corresponding array of labels $L_{r,i}$. The persons were regular users of point cloud data having different levels of expertise (2 PhDs, 4 PhD candidates and 4 bachelor students) in point cloud processing and visualization. The two rounds of labelling were spaced of one month. No pre-classification was used for any of the rounds. The labelling was conducted in the environment of Cloud Compare software with the segmentation tool, leaving the display options (pixel size, colour, and use of lighting models) up to each person to decide. The points were labelled into nine classes: road, sidewalk, curb, building, car, vegetation, pole-like objects (traffic lights, signs, and streetlights), furniture, and others. The following criteria were established to ensure consistency between labels from different

**Table 1**
Technical specifications of Riegl VUX-1HA scanner.

| Field of view (vertical/horizontal) | 360° full circle |
|---|---|
| Angular resolution (vertical/horizontal) | 0.001° |
| Range (m) | 1.2–420 |
| Accuracy (mm) | 5 |
| Pts/s | Up to 1,000,000 |
| Wavelength (nm) | Near infrared |

persons:

- All points behind the facades were considered as *building* points, including floors of entrances and indoor spaces.
- Only unmoving cars were considered as *cars*. Motorbikes, large trucks, and any other vehicle in motion were defined as others.
- Traffic lights, sign and streetlight posts were considered as pole-like objects. But the bollard points were considered as others.
- Ramps were classified within *sidewalk* points.
- Any other points and that were from unrecognizable objects were classified as others.

To obtain the ground truth, the point cloud was opened with CloudCompare, and using the segmentation tool the points of each class were obtained. Some visualization tools in CloudCompare software were used to segment each class easily. Different views, size of the points or colour of the point cloud. Nine point clouds were obtained in the first step of the manual labelling process, one for each class, Fig. 1. Then, these point clouds were merged, generating a new index.

The time used for manual labelling of each point cloud was approximately 30 min per person and round. One labelled point cloud is seen in Fig. 2. Subsequently, label arrays were concatenated to the input point cloud $P(X, Y, Z, L_{1,1} \cdots L_{1,10}, L_{2,1} \cdots L_{2,10})$, being $L_{r,i}$ the label, where $r$ is the number of round ($r = 1 \text{and} 2$) and $i$ is the person who did the labelling ($i = 1, 2, \cdots, 10$).

### 3.3. Discordance

The issue of discordance measurement between labels can be presented as a classification problem of confusion. However, unlike the conventional problem, in this case there was no reference data since the ground truth was also subject to interpretation. It was also not a one-to-one comparison, since several sets of labels were compared. Therefore, it is proposed to use a triangular matrix $M(9 \times 9)$ classes whose main diagonal was composed of the points that had the same label and the other positions were discordances. Concordance and discordance were defined as follows:

- Concordance means that the same class assignment to one point by all persons in all rounds $\forall L_n = L_{n+1}$. In such a case, the point is counted on the main diagonal of the corresponding class.
- Discordance is defined as the existence of different classes assigned to the same point $\exists L_n \neq L_{n+1}$. It is counted as discordance between the corresponding confusions in the matrix.

To establish a relation between both, a discordance ratio was defined ($r_{dc}$) that accounts for the discordances with respect to the concordance of each point as in Equation (1), where $d$ is discordance and $c$ concordance in $r_{dc}$.

$$r_{dc} = \frac{discordances}{pointsinconcordance} \tag{1}$$

### 3.4. Significance of test statistic

The null hypothesis ($H_0$) was defined to test if there are any significant difference of labelling from the same person within two different rounds and from 10 different persons within the same class in the same round. The well-known Marascuillo proportion test procedure was employed to test the $H_0$. It enables to simultaneously test the differences of all pairs of proportions when there are several populations under investigation (NIST, 2022). The first step of this procedure is to compute the proportions per class of each person, $p_i = l_{class}/l$ , being $l_{class}$ the number of points of each class, $l$ the total number of points and ($i = 1, 2, \cdots, 10$). Then, the differences of all pairs of proportions, $p_m - p_n$, are calculated where $m$ and $n$ represent two different persons who did the classification, among all $j(j-1)/2$ pairs of proportions, being $j = 10$. In total, there were 45 combinations when considering 10 people. The corresponding critical values for the Marascuillo procedure were obtained following the Equation (2), where $p_m$ is the proportion value of one person and $p_n$ the proportion value of another person. $\chi^2_{1-\alpha,j-1}$ is the chi-square value and $l_m$ and $l_n$ the total number of labelled points by the corresponding person.

$$r_{mn} = \sqrt{\chi^2_{1-\alpha,j-1}} \sqrt{\frac{p_m(1-p_m)}{l_m} + \frac{p_n(1-p_n)}{l_n}} \tag{2}$$

The chi-square value was obtained from a table predefined. To achieve this value the degrees of freedom (DOF) were calculated, in that case $j-1 = 9$. Then the significance level $\alpha$ was chosen. Often, researchers choose a significant level $\alpha$ equal to 0.01, 0.05 or 0.1 (Berman, 2022), therefore, significance was calculated with these three $\alpha$ values. The last step was to compare the $j(j-1)/2$ test statistics against its corresponding critical $r_{mn}$ value. Those pairs that had an estimate of the test statistic exceeds the critical value were significant at the $\alpha$ (what is the value 0.01,0.05 or 0.1) level.

The computational complexity of the Marascuillo equation (Equation (2) depends mainly on the input values, which are the sample sizes, the proportions of the groups and the significance level $\alpha$. The Marascuillo equation can be expressed in terms of polynomial time P according to the input size. Since the Marascuillo equation involves performing basic arithmetic operations (addition, multiplication, division and square roots), all these operations can be performed in constant time. In addition, the calculation of the chi-square value can also be performed in polynomial time. Therefore, the time complexity of the Marascuillo equation can be considered polynomial as a function of the input size, which means that the execution time increases polynomially as the data size increases. In terms of computational complexity notation, the Marascuillo equation has a complexity of O(n^2), where n is the number of groups being compared. It is important to note that although the Marascuillo equation can be expressed in terms of polynomial time, this is only true in terms of time complexity, and does not consider other factors that may affect computational performance, such as the amount of memory required to store the data or the number of input/output operations that must be performed.

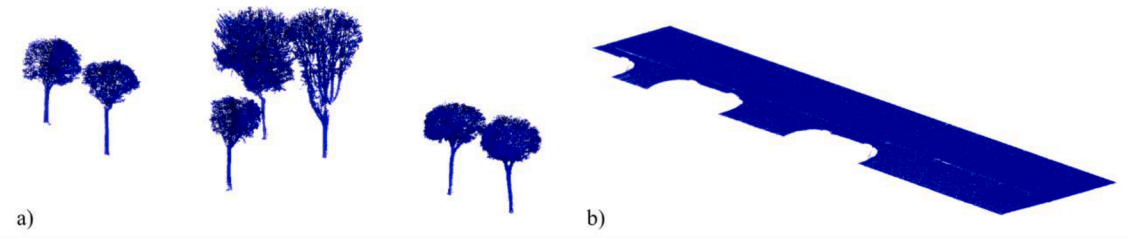


**Fig. 1.** Preliminary point clouds of manual classification: a) *vegetation* class, and b) *road* class.

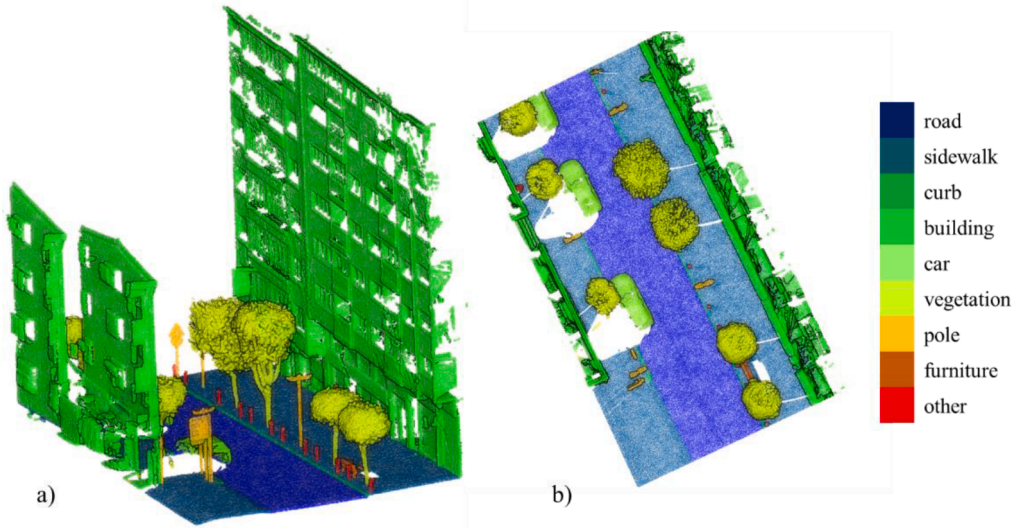


**Fig. 2.** Point cloud labelled and coloured by different classes: a) perspective view, and b) top view.

## 4. Results and analysis

### 4.1. Counting and localization of discordances

Table 2 groups points according to the number of different assigned labels they present. The values were calculated by counting the number of different classes associated with each point per round (first and second) as well as together (total), considering in the same calculous the labelled point clouds in both rounds. Considering two or more classes assigned to the same point, in the first round, 100 thousand discordances were counted and in the second round, 75 thousand discordances. From the joint analysis it can be deduced that part of the discordances existing in the second round already appeared in the first, although new ones also appeared. The number of discordances represented 2.6% points in the cloud. It is striking how 19 points in the whole process were labelled as five different classes (*sidewalk, car, vegetation, pole*, and *others*) out of nine possible classes. Fig. 3 shows discordances in the labelling for the points that are highlighted in red. Discordances were in adjacency zones between elements although varying in extension, so that the most affected elements are those with larger perimeters, especially in relation to the number of points.

Table 3 shows the points with labelled concordance and discordance and the percentage of discordance points (Equation (3). There were few differences between rounds in most of classes. Only for the classes *curbs*, *furniture* and *others,* a difference of 1% is exceeded between the two rounds. In particular, the class *others* doubled the percentage of discordance between both rounds due to only one object. Generally, these three classes had the highest percentage of discordance, followed by *sidewalks* and *pole-like* objects.

$$Discordance(\%) = \frac{no.pointsindiscordance}{no.pointsinconcordance + no.pointsindiscordance} \times 100 \quad (3)$$

**Table 2**
Points according to the number of classes assigned.

| | | Round 1 | Round 2 | Total |
|---|---|---|---|---|
| number of classes | 1 | 3,980,139 | 4,005,400 | 3,964,981 |
| | 2 | 97,383 | 68,421 | 106,346 |
| | 3 | 2,877 | 6,381 | 8,793 |
| | 4 | 30 | 227 | 290 |
| | 5 | 0 | 0 | 19 |
| | 6 | 0 | 0 | 0 |

Discordances between classes are compiled in Table 4. Mainly and almost exclusively, discordances occurred between adjacent elements. However, some noise points were accounted for between non-adjacent elements, such as *road-buildings*, or *road-poles*. *Sidewalks* was the class that showed discordances to more different classes. *Sidewalk* discordances were produced towards the rest of the classes, since the sidewalks were adjacent to all of them, both for limiting the contour of road, curbs and buildings, and for being the ground element on which vegetation, poles, furniture and other objects are located. In terms of quantity, discordances from *curbs* to *sidewalk* and *road* classes were very important. In both, the number of discordances exceeded the number of matching points. Table 5 shows the number of points with labelled concordance, associated discordances and $r_{dc}$. *Curb* class, with more than 3 discordances per concordance point, followed by the class *other*, were the elements with the highest $r_{dc}$. Fig. 4 shows an enlargement of a curb. In the upper adjacent sidewalk area, points with discordance in the labelling were detected with a width of 19 cm, while the curb size is only 12 cm. To this must be added the 8 cm wide discordant points with the road. In the rest of the classes, $r_{dc}$ was significantly lower.

### 4.2. Significance of discordances

To measure the significance of discordances the Marascuillo proportion test procedure was employed to test the $H_0$. The difference of all pairs of proportion were calculated and compared with the critical value, which was obtained following the Equation (2) and considering three different $\alpha$ values (0.01, 0.05 and 0.1). Discordances in classification were significative if $p_m - p_n > r_{mn}$. Table 6 shows the percentage significance of the point classification between people in the first round, using $\alpha = 0.01$, 0.05 and 0.1. The values obtained using different $\alpha$ are almost the same, just values of *sidewalk*, *curb*, *building* and *others* classes, were slightly different between using $\alpha = 0.01$ and $\alpha = 0.05$ or $\alpha = 0.1$. The classes *vegetation*, *pole-like* objects and *furniture* resulted no significant differences between the discordances from different people in the same round. Class *vehicle* presented a significant percentage of 20%, caused only by the significance of one person with respect to the others. *Road*, *sidewalk*, *curb*, *building* and *others* classes presented higher percentages of significance. These significances are due to adjacent elements, which were classified with different classes by each person. Table 7 exemplifies the significance results ($\alpha = 0.05$) of *road*, *sidewalk*, and *curb* classes, comparing the person $I = 1$ with the rest of the people.

Table 8 shows the percentage of significative labelling differences between two rounds of classification by the same person, using $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Results show that there are no differences

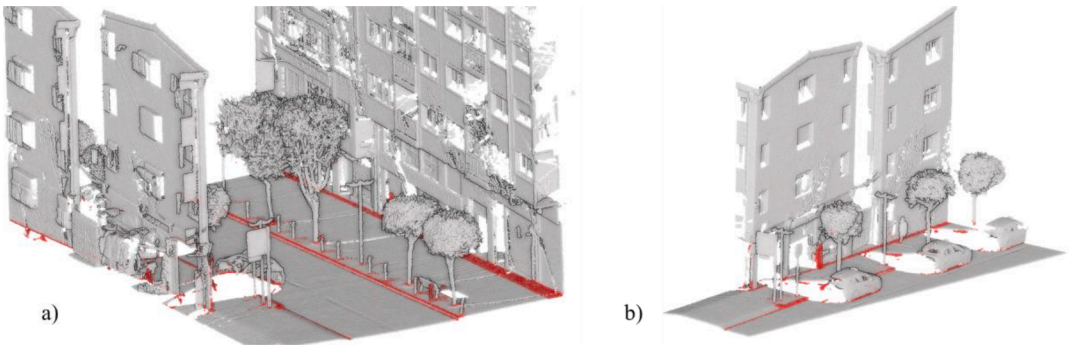


**Fig. 3.** Point cloud with discordances coloured in red.

**Table 3**
Counting the number of points in labelled concordance and points in labelled discordance.

| | First round | | | Second round | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Conc. | Disc. | Disc.(%) | Conc. | Disc. | Disc.(%) | Conc. | Disc. | Disc.(%) |
| Road | 803,810 | 485 | 0.1% | 820,226 | 513 | 0.1% | 803,569 | 1,440 | 0.2% |
| Sidewalk | 562,776 | 2,698 | 0.5% | 576,278 | 6,166 | 1.1% | 558,811 | 8,395 | 1.5% |
| Curb | 174,40 | 551 | 3.1% | 18,627 | 121 | 0.6% | 15,654 | 989 | 5.9% |
| Buildings | 1,939,196 | 1,570 | 0.1% | 1,932,733 | 4,164 | 0.2% | 1,932,328 | 5,066 | 0.3% |
| Car | 157,524 | 196 | 0.1% | 156,925 | 623 | 0.4% | 156,809 | 887 | 0.6% |
| Vegetation | 445,097 | 571 | 0.1% | 445,314 | 1,484 | 0.3% | 444,813 | 1,737 | 0.4% |
| Pole-like | 38,677 | 253 | 0.6% | 38,570 | 293 | 0.8% | 38,325 | 646 | 1.7% |
| Furniture | 5,061 | 23 | 0.5% | 5,060 | 268 | 5.0% | 5,014 | 314 | 5.9% |
| Others | 10,558 | 2,404 | 18.5% | 11,667 | 6,419 | 35.5% | 9,658 | 8,160 | 45.8% |

**Table 4**
Confusion matrix of points in two rounds of manual labelling (Side.: Sidewalk, Build.: Building, Veg.: Vegetation, Furn.: Furniture).

| Road | Side. | Curb | Build. | Car | Veg. | Pole | Furn. | Others | |
|---|---|---|---|---|---|---|---|---|---|
| 803,569 | 4,554 | 20,231 | 1 | 8,934 | 12 | 5 | 0 | 746 | **Road** |
| | 558,811 | 29,772 | 37,071 | 189 | 3,939 | 1,440 | 576 | 11,691 | **Side.** |
| | | 15,654 | 2 | 22 | 3 | 97 | 35 | 140 | **Curb** |
| | | | 1,932,328 | 0 | 329 | 368 | 0 | 9,081 | **Build.** |
| | | | | 156,809 | 233 | 42 | 0 | 1,358 | **Car** |
| | | | | | 444,813 | 47 | 0 | 1,906 | **Veg.** |
| | | | | | | 38,325 | 0 | 758 | **Pole** |
| | | | | | | | 5,014 | 1,073 | **Furn.** |
| | | | | | | | | 9,658 | **others** |

**Table 5**
Manually labelled points of concordance (conc.), discordances (disc.) and $r_{dc}$ (Side.: Sidewalk, Build.: Building, Veg.: Vegetation, Furn.: Furniture).

| | First round | | | Second round | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | Conc. | Disc. | $r_{dc}$ | Conc. | Disc. | $r_{dc}$ | Conc. | Disc. | $r_{dc}$ |
| Road | 803,810 | 30,691 | 0.038 | 820,226 | 13,857 | 0.017 | 803,569 | 34,483 | 0.043 |
| Side. | 562,776 | 71,989 | 0.128 | 576,278 | 65,648 | 0.114 | 558,811 | 89,232 | 0.160 |
| Curb | 17,440 | 46,936 | 2.691 | 18,627 | 33,674 | 1.808 | 15,654 | 50,302 | 3.213 |
| Build | 1,939,196 | 35,923 | 0.019 | 1,932,733 | 31,515 | 0.016 | 1,932,328 | 46,852 | 0.024 |
| Car | 157,524 | 8,974 | 0.057 | 156,925 | 3,821 | 0.024 | 156,809 | 10,778 | 0.069 |
| Veg. | 445,097 | 4,037 | 0.009 | 445,314 | 4,719 | 0.011 | 444,813 | 6,469 | 0.015 |
| Pole | 38,677 | 1,725 | 0.045 | 38,570 | 1,619 | 0.042 | 38,325 | 2,757 | 0.072 |
| Furn. | 5,061 | 394 | 0.078 | 5,060 | 1,494 | 0.295 | 5,014 | 1,684 | 0.336 |
| Others | 10,558 | 11,719 | 1.110 | 11,667 | 21,505 | 1.843 | 9,658 | 26,753 | 2.770 |

using the three values of $\alpha$ in class *vehicle*, *vegetation*, *pole-like*, *furniture* and *others*. *Vegetation* and *pole-like* class do not present significative changes which means that people classified these classes nearly the same in both rounds. *Vehicle* and *furniture* classes present a 10% of significance, therefore there was a person who made significant changes in the classification of these classes between rounds. Significance percentages of *sidewalk* and *curb* classes are higher. Moreover, class *curb* presents a 100% of significance. Table 9 exemplifies the significance results ($\alpha = 0.05$) of two rounds done by same persons. Regarding the class *sidewalk* more than the half of the people did significative changes between rounds, and in the *curb* class all the people did significative changes between rounds.

## 5. Discussion

Point clouds are manually labelled using CloudCompare (CloudCompare, 2023), which is a free software for point cloud processing. Although, there are more software that can be used to obtain the labelled

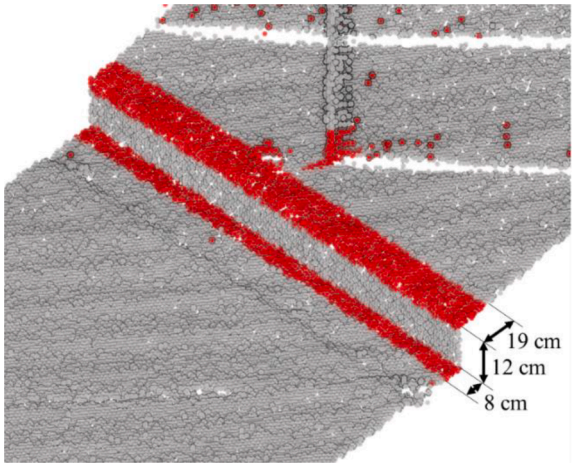


**Fig. 4.** Zoom to the curb area (points with labels in discordance highlighted in red).

point clouds. Amazon (AWS, 2023), Matlab (MathWorks, 2023) or Segments.ai (Segments.ai, 2023) offers software to manually label point clouds, however they are not free software. The generated data from CloudCompare is used to analyse differences between labelling by different people and rounds, generating a ground truth. All the points classified with the same class by all the people and rounds are considered as ground truth. The points which are labelled with different classes by people and rounds are analysed and a voting system is applied to obtain the ground truth of that points.

Ground truth is also obtained manually by other authors. Several existing datasets classified point clouds by hand using CloudCompare, (Deschaud et al., 2021; Hu et al., 2021; Roynard et al., 2018). However, these works do not compare the manual labelling done by several people and do not measure the discordances between the different classifications. Given the amount of data of these studies, it is assumed that several people could have been involved in the manual labelling process, although this information is not mentioned. While the manual labelling is done by the same person, in the present work was reflected that the same person also obtains discordances in the manual labelling. Table 10 shows the percentage of discordance points (Equation (3) obtained in some classes, performing the manual labelling in two rounds by the

**Table 6**
Percentage of significative labelling differences between people in the same round.

| $\alpha$ | Road | Sidewalk | Curb | Building | Vehicle | Vegetation | Pole-like | Furniture | Others |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 44.4% | 48.9% | 80.0% | 35.6% | 20.0% | 0.0% | 0.0% | 0.0% | 64.4% |
| 0.05 | 44.4% | 62.2% | 86.7% | 48.9% | 20.0% | 0.0% | 0.0% | 0.0% | 66.7% |
| 0.1 | 44.4% | 62.2% | 88.9% | 48.9% | 20.0% | 0.0% | 0.0% | 0.0% | 66.7% |

**Table 7**
Comparison between person $i = 1$ and remaining people ($\alpha = 0.05$).

| Road | | | | Sidewalk | | | Curb | | |
|---|---|---|---|---|---|---|---|---|---|
| | $p_m - p_n$ | $r_{mn}$ | **Significative** | $p_m - p_n$ | $r_{mn}$ | **Significative** | $p_m - p_n$ | $r_{mn}$ | **Significative** |
| $L_{1,1}$-$L_{1,2}$ | 0.0019 | 0.0012 | Yes | 0.0017 | 0.001 | Yes | 0.00033 | 0.00022 | Yes |
| $L_{1,1}$-$L_{1,3}$ | 0.0017 | 0.0012 | Yes | 0.0011 | 0.001 | Yes | 0.00094 | 0.00023 | Yes |
| $L_{1,1}$-$L_{1,4}$ | 0.0011 | 0.0012 | No | 0.0005 | 0.001 | No | 0.0011 | 0.00023 | Yes |
| $L_{1,1}$-$L_{1,5}$ | 0.0009 | 0.0012 | No | 0.0091 | 0.001 | Yes | 0.00771 | 0.00029 | Yes |
| $L_{1,1}$-$L_{1,6}$ | 0.0018 | 0.0012 | Yes | 0.0014 | 0.001 | Yes | 0.00044 | 0.00023 | Yes |
| $L_{1,1}$-$L_{1,7}$ | 0.0006 | 0.0012 | No | 0.0061 | 0.001 | Yes | 0.00737 | 0.00028 | Yes |
| $L_{1,1}$-$L_{1,8}$ | 0.0017 | 0.0012 | Yes | 0.0002 | 0.001 | No | 0.00003 | 0.00022 | No |
| $L_{1,1}$-$L_{1,9}$ | 0.0016 | 0.0012 | Yes | 0.0013 | 0.001 | Yes | 0.00002 | 0.00022 | No |
| $L_{1,1}$-$L_{1,10}$ | 0.0016 | 0.0012 | yes | 0.0017 | 0.001 | Yes | 0.00025 | 0.00023 | Yes |

**Table 8**
Percentage of significative labelling differences between rounds of the same people.

| $\alpha$ | Road | Sidewalk | Curb | Building | Vehicle | Vegetation | Pole-like | Furniture | Others |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 20% | 50% | 90% | 20% | 10% | 0% | 0% | 10% | 40% |
| 0.05 | 40% | 60% | 100% | 40% | 10% | 0% | 0% | 10% | 40% |
| 0.1 | 40% | 60% | 100% | 40% | 10% | 0% | 0% | 10% | 40% |

**Table 9**
Comparison between rounds of the same person ($\alpha = 0.05$).

| Road | | | | Sidewalk | | | Curb | | |
|---|---|---|---|---|---|---|---|---|---|
| | $p_m - p_n$ | $r_{mn}$ | **Significative** | $p_m - p_n$ | $r_{mn}$ | **Significative** | $p_m - p_n$ | $r_{mn}$ | **Significative** |
| $L_{1,1}$-$L_{2,1}$ | 0.00146 | 0.00116 | Yes | 0.0021 | 0.001 | Yes | 0.0005 | 0.0002 | No |
| $L_{1,2}$-$L_{2,2}$ | 0.00017 | 0.00116 | No | 0.0003 | 0.001 | No | 0.0004 | 0.0002 | Yes |
| $L_{1,3}$-$L_{2,3}$ | 0.00066 | 0.00116 | No | 0.0014 | 0.001 | Yes | 0.0005 | 0.0002 | Yes |
| $L_{1,4}$-$L_{2,4}$ | 0.00122 | 0.00116 | Yes | 0.0005 | 0.001 | No | 0.0008 | 0.0002 | No |
| $L_{1,5}$-$L_{2,5}$ | 0.00152 | 0.00116 | Yes | 0.0073 | 0.001 | Yes | 0.0013 | 0.0003 | Yes |
| $L_{1,6}$-$L_{2,6}$ | 0.00033 | 0.00116 | No | 0.0005 | 0.001 | No | 0.0008 | 0.0002 | No |
| $L_{1,7}$-$L_{2,7}$ | 0.0024 | 0.00116 | Yes | 0.0074 | 0.001 | Yes | 0.0061 | 0.0003 | No |
| $L_{1,8}$-$L_{2,8}$ | 0.00067 | 0.00116 | No | 0.0013 | 0.001 | Yes | 0.0011 | 0.0002 | No |
| $L_{1,9}$-$L_{2,9}$ | 0.0001 | 0.00116 | No | 0.0014 | 0.001 | Yes | 0.0004 | 0.0002 | Yes |
| $L_{1,10}$-$L_{2,10}$ | 0.00003 | 0.00116 | No | 0.0004 | 0.001 | No | 0.0002 | 0.0002 | No |

**Table 10**
Discordances of the same person between rounds.

| | Road | Sidewalk | Curb | Building |
|---|---|---|---|---|
| $L_{1,2}$-$L_{2,2}$ | 0,085 | 0,186 | 6,604 | 0,274 |
| $L_{1,3}$-$L_{2,3}$ | 0,325 | 0,890 | 6,580 | 0,170 |
| $L_{1,9}$-$L_{2,9}$ | 0,051 | 0,895 | 6,688 | 0,136 |

same person.

In the present method the discordances are measured considering manual labelling of 10 people in two rounds, being the novelty of this work. Most of the discordances are produced in the unions between road-curb-sidewalk. Also, in the union between sidewalks and buildings, and in the areas of the sidewalk and road where there are street furniture, vegetation and cars. In these areas discordances are produced in the union of these objects with the sidewalk and road in the case of cars.

The relation between labelled discordances and the balance between classes is analysed to explain the results of DL methods. Since Neural Networks (NNs) are black boxes, only inputs and outputs can be analysed to find possible improvements in their performance. The mainly used inputs to explain the NNs are the number of samples and the balance between classes, but in semantic segmentation, it is very difficult to obtain a balanced number of samples because the urban environment (and almost in every scene) never contains the same number of points per class that defines objects of different sizes.

Table 11 compiles the number of points used in the training process and the results of the PointNet++ (Qi et al., 2017b) application trained with those points. More information on the survey, labelling and training process can be found in González-Collazo et al. (2022). Table 11 also compiles the $r_{dc}$ and percentage of points with concordance labelling per class presented in this paper. Table 12 shows the correlation coefficient $R^2$ (Equation (4) calculated to look for relation between the number of samples, $r_{dc}$ and percentage of concordance and results from PointNet++ predictions in terms of precision, recall, *f*-score and Intersection over Union (IoU). The $R^2$ between number of points and results was quite low. The $r_{dc}$ proposed in this work improved the correlation but the percentage of concordance showed the highest correlation, reaching an $R^2 = 0.765$ with the *f*-score, so it can be deduced that concordances and discordances in labelling were conclusive factors in the training process.

$$R^2 = \frac{\sum_{t=1}^{T}(\widehat{A}_t - \overline{A})^2}{\sum_{t=1}^{T}(A_t - \overline{A})^2} \tag{4}$$

The present method was tested with PointNet++, however, there are more convolutional neural networks. KPConv (Thomas et al., 2019) is a convolutional neural network for processing 3D point clouds. KPConv is based on PointNet++ and uses an adaptative convolution technique to capture features at different scales. ShellNet (Zhang et al., 2019) is a convolutional neural network used for the classification of objects in 3D point clouds. ShellNet uses a similar architecture to PointNet++, but also incorporates a shell projection technique to capture features of objects. ParNet (Mo et al., 2018), RandLA-Net (Hu et al., 2019) and Point ASNL (Yan et al., 2020) are convolutional neural networks used for semantic segmentation of 3D point clouds. ParNet is based on the PointNet++ architecture and uses hierarchical clustering technique to capture features at different scales. RandLA-Net is based on the PointNet++ architecture and uses random clustering technique to capture local features. PoinASNL is based on the PointNet++ architecture and uses a selective clustering technique to capture local features. Although there are more current convolutional neural networks, it should be noted that many of these neural networks are based on the PointNet++ architecture due to its efficiency in 3D point cloud processing.

**Table 12**
Correlation coefficient $R^2$ between number of training points, $r_{dc}$, percentage of concordance points and PointNet++ outcome metrics.

| | precision | recall | *f*-score | IoU |
|---|---|---|---|---|
| no. points train | 0.373 | 0.245 | 0.309 | 0.426 |
| $r_{dc}$ | 0.443 | 0.429 | 0.486 | 0.475 |
| % conc. | 0.672 | 0.359 | 0.765 | 0.624 |

In the best of the authors knowledge, no DL-based methods were found with results per class that reached the percentage of concordant points accounted in this work, although some are beginning to come close. Good labelling is fundamental for Deep Learning algorithms, but it is very difficult to achieve. Errors aside, which were not the object of study in this work, the same dataset labelled by one or several people presents significant differences between elements and areas, which leads to a change in the geometric-topological features extracted by NNs. A solution to manual labelling discordances may be the use of simulators to generate automatically labelled data, such as CARLA (Dosovitskiy et al., 2017) or NVIDIA Omniverse (Hummel & van Kooten, 2019). However, it should also be evaluated the realism of such synthetic data with respect to the quality of real data or to unforeseen situations in complex built environments.

## 6. Conclusion

This research presents the first study of the discordances in manual labelling of MLS point clouds. To measure the discordances, an urban MLS point cloud of 30 m road-length was labelled twice (2 rounds) by 10 people. A ratio test was defined to relate discordances and concordance in labelling. Discordance and the significance of discordances in manual labelling between individuals and rounds were also measured.

The results showed that, although most of the points were classified accordingly as the same class by all the participants, there are several points that were classified differently into two (or more) classes. The percentage of discordance was lower than 1% in almost all classes, however, in classes such as curbs, there were 5.9% of discordant points and 3.2 discordances for each point with concordance by all persons. The significance of discordance of the labelled data was measured, obtaining the higher significant differences between people and rounds in classes *sidewalk* and *curb*. The high significance and discordances of sidewalks are because they are adjacent with almost all other classes (*roads, curb, building, poles, trees, furniture,* and *others*). Discordances in

**Table 11**
Values per class of number of training/validation points, $r_{dc}$, percentage of concordance points and PointNet++ outcome metrics (Side.: Sidewalk, Build.: Building, Veg.: Vegetation, Furn.: Furniture).

| | no. points train | $r_{dc}$ | % conc. | precision | recall | *f*-score | IoU |
|---|---|---|---|---|---|---|---|
| Road | 44,698,817 | 0.043 | 99.8% | 0.844 | 0.984 | 0.908 | 0.832 |
| Side. | 24,129,907 | 0.160 | 98.5% | 0.726 | 0.789 | 0.756 | 0.608 |
| Curb | 1,813,755 | 3.213 | 94.1% | 0.576 | 0.724 | 0.641 | 0.472 |
| Build. | 135,957,248 | 0.024 | 99.7% | 0.996 | 0.907 | 0.950 | 0.904 |
| Car | 7,692,577 | 0.069 | 99.4% | 0.885 | 0.784 | 0.831 | 0.711 |
| Veg. | 9,454,048 | 0.015 | 99.6% | 0.702 | 0.963 | 0.812 | 0.683 |
| Pole | 1,068,977 | 0.072 | 98.3% | 0.466 | 0.737 | 0.571 | 0.400 |
| Others | 4,877,732 | 2.770 | 54.2% | 0.141 | 0.662 | 0.233 | 0.132 |

*curbs*, are due to the length and narrowness, as well as the difficulty in delimit a common border.

In addition, PointNet++ (González-Collazo et al., 2022) based semantic segmentation results were analysed, we found a relationship ($R^2$ = 0.765) between the *F-score* and the percentage of points labelled in concordance, which clearly explains the causes of the misclassification than the balanced number of points between the classes. Future work will extend the measurements to other types of laser scanning data and in different scenarios. Also, the proposed method will be applied to others deep learning architectures, to analyse how manual labelling affects in their results.

## CRediT authorship contribution statement

**Silvia María González-Collazo:** Methodology, Software, Formal analysis, Writing – original draft. **Jesús Balado:** Conceptualization, Methodology, Software, Writing – review & editing. **Elena González:** Writing – review & editing. **Abdul Nurunnabi:** Conceptualization, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

The authors would like to thank the point cloud labelling volunteers for their time and effort. This research has received funding from Xunta de Galicia through human resources grant (ED481B-2019-061) and from the Government of Spain through project PID2019-105221RB-C43 funded by MCIN/AEI/10.13039/501100011033. Mr Nurunnabi is with the Project 2019-05-030-24, SOLSTICE - Programme Fonds Européen de Developpment Régional (FEDER)/Ministère de l'Economie of the G. D. of Luxembourg.This paper was carried out in the framework of the InfraROB project (Maintaining integrity, performance and safety of the road infrastructure through autonomous robotized solutions and modularization), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 955337. It reflects only the authors' views. Neither the European Climate, Infrastructure, and Environment Executive Agency (CINEA) nor the European Commission is in any way responsible for any use that may be made of the information it contains. Authors also would like to thank to CESGA for the use of their servers. Funding for open access charge: Universidade de Vigo/CISUG.

## References

Abdelaziz Ismael, S. A., Mohammed, A., & Hefny, H. (2020). An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artificial Intelligence in Medicine, 102*, Article 101779. https://doi.org/10.1016/j.artmed.2019.101779.

Armeni, I., Sax, A., Zamir, A.~R, & Savarese, S. (2017). *Joint 2D–3D-Semantic Data for Indoor Scene Understanding*. ArXiv E-Prints.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., & Savarese, S. (2016). 3D Semantic Parsing of Large-Scale Indoor Spaces. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 1534–1543. https://doi.org/10.1109/CVPR.2016.170

AWS. (2023). *AWS*. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-point-cloud.html.

Balado, J., Martínez-Sánchez, J., Arias, P., & Novo, A. (2019). Road environment semantic segmentation with deep learning from mls point cloud data. *Sensors (Switzerland), 19*(16). https://doi.org/10.3390/s19163466

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., & Gall, J. (2019). SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. *Proceedings of the IEEE International Conference on Computer Vision*, 9297–9307.

Berman, H. . (2022). *Chi-Square Test of Independence*. "Chi-Square Test of Independence." https://stattrek.com/chi-square-test/independence.

Boulch, A. (2020). ConvPoint: Continuous convolutions for point cloud processing. *Computers & Graphics, 88*, 24–34. https://doi.org/10.1016/j.cag.2020.02.005.

Charles, R., Su, H., Mo, K., & Guibas, L. (2017). *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. 77–85. https://doi.org/10.1109/CVPR.2017.16.

Choy, C., Gwak, J., & Savarese, S. (2019). 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chun, L., Doudou, Z., Akram, A., Hangbin, W., Shoujun, J., Zeran, X., & Han, Y. (2021). *Tongji-3D-Dataset*. https://github.com/ZivKidd/Tongji-3D-Dataset.

CloudCompare. (2023). *CloudCompare*. https://www.cloudcompare.org/main.html.

Deschaud, J.-E., Duque, D., Richa, J. P., Velasco-Forero, S., Marcotegui, B., & Goulette, F. (2021). Paris-CARLA-3D: A Real and Synthetic Outdoor Point Cloud Dataset for Challenging Tasks in 3D Mapping. *Remote Sensing, 13*(22). https://doi.org/10.3390/rs13224713

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. In S. Levine, V. Vanhoucke, & K. Goldberg (Eds.), *Proceedings of the 1st Annual Conference on Robot Learning* (Vol. 78, pp. 1–16). PMLR.

González-Collazo, S. M., Balado, J., Garrido, I., Grandío, J., Rashdi, R., Tsiranidou, E., del Río-Barral, P., Rúa, E., Puente, I., & Lorenzo, H. (2022). Santiago Urban Dataset Sud: Combination of Handled and Mobile Laser Scanning Point Clouds. *SSRN*.

González, E., Balado, J., Arias, P., & Lorenzo, H. (2022). Realistic correction of sky-coloured points in Mobile Laser Scanning point clouds. *Optics & Laser Technology, 149*, Article 107807. https://doi.org/10.1016/j.optlastec.2021.107807.

Guiotte, F., Pham, M.-T., Dambreville, R., Corpetti, T., & Lefèvre, S. (2020). Semantic Segmentation of LiDAR Points Clouds: Rasterization Beyond Digital Elevation Models. *IEEE Geoscience and Remote Sensing Letters, 17*(11), 2016–2019. https://doi.org/10.1109/LGRS.2019.2958858

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., & Pollefeys, M. (2017). Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark. *ArXiv*, *abs/1704.0*.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., & Zhang, B. (2021). More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing, 59*(5), 4340–4354. https://doi.org/10.1109/tgrs.2020.3016820

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., & Chanussot, J. (2022). SpectralFormer}: Rethinking Hyperspectral Image Classification With Transformers. *{IEEE} Transactions on Geoscience and Remote Sensing, 60*, 1–15. https://doi.org/10.1109/tgrs.2021.3130716

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., & Markham, A. (2021). Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A Dataset, Benchmarks and Challenges. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021*, 4975–4985. https://doi.org/10.1109/CVPR46437.2021.00494

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., & Markham, A. (2019). *RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds (CVPR 2020 Oral)*.

Hummel, M., & van Kooten, K. (2019). *Leveraging NVIDIA Omniverse for In Situ Visualization* (M. Weiland, G. Juckeland, S. Alam, & H. Jagode (eds.); pp. 634–642). Springer International Publishing.

Landrieu, L., & Simonovsky, M. (2018). Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018*, 4558–4567. https://doi.org/10.1109/CVPR.2018.00479

Lê, H.-Â., Guiotte, F., Pham, M.-T., Lefèvre, S., & Corpetti, T. (2022). Learning Digital Terrain Models From Point Clouds: ALS2DTM Dataset and Rasterization-Based GAN. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15*, 4980–4989. https://doi.org/10.1109/JSTARS.2022.3182030

Li, M., Xie, Y., Shen, Y., Ke, B., Qiao, R., Ren, B., Lin, S., & Ma, L. (2022). HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Contrastive Regularization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14930–14939.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., & Benediktsson, J. A. (2019). Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Transactions on Geoscience and Remote Sensing, 57*(9), 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932

Liu, C., Zeng, D., Akbar, A., Wu, H., Jia, S., Xu, Z., & Yue, H. (2022). Context-Aware Network for Semantic Segmentation Toward Large-Scale Point Clouds in Urban Environments. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–15. https://doi.org/10.1109/TGRS.2022.3182776

Lu, T., Wang, L., & Wu, G. (2021). CGA-Net: Category Guided Aggregation for Point Cloud Semantic Segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021*, 11688–11697. https://doi.org/10.1109/CVPR46437.2021.01152

Ma, L., Li, Y., Li, J., Wang, C., Wang, R., & Chapman, M. A. (2018). Mobile Laser Scanned Point-Clouds for Road Object Detection and Extraction: A Review. *Remote Sensing, 10*(10). https://doi.org/10.3390/rs10101531

MathWorks. (2023). *MathWorks*. https://es.mathworks.com/help/driving/ug/label-lidar-point-clouds-for-object-detection.html.

Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., & Su, H. (2018). PartNet: {A} Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding. *CoRR, abs/1812.0*. http://arxiv.org/abs/1812.02713.

NIST. (2022). *Comparing multiple proportions: The Marascuillo procedure*. https://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm.

Nurunnabi, A., Teferle, N., Li, J., Lindenbergh, R., & Hunegnaw, A. (2021). An Efficient Deep Learning Approach for Ground Point Filtering in Aerial Laser Scanning Point Clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B1-2*. https://doi.org/10.5194/isprs-archives-XLIII-B1-2021-31-2021.

Nurunnabi, A., Teferle, N., Li, J., Lindenbergh, R., & Parvaz, S. (2021). *Investigation of PointNet for Semantic Segmentation of Large-Scale Outdoor Point Clouds*.

Paz Mouriño, S. de, Balado, J., & Arias, P. (2021). Multiview Rasterization of Street Cross-sections Acquired with Mobile Laser Scanning for Semantic Segmentation with Convolutional Neural Networks. *IEEE EUROCON 2021 - 19th International Conference on Smart Technologies*, 35–39. https://doi.org/10.1109/EUROCON52738.2021.9535645.

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., … Lingua, A. M. (2020). Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage. *Remote Sensing, 12*(6). https://doi.org/10.3390/rs12061005

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017a). PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *CoRR*, *abs/1706.0*.

Richter, R., & Döllner, J. (2014). Concepts and techniques for integration, analysis and visualization of massive 3D point clouds. *Computers, Environment and Urban Systems, 45*, 114–124. https://doi.org/10.1016/j.compenvurbsys.2013.07.004.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 3234–3243. https://doi.org/10.1109/CVPR.2016.352

Roynard, X., Deschaud, J.-E., & Goulette, F. (2018). Paris-Lille-3D: A Point Cloud Dataset for Urban Scene Segmentation and Classification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018*, 2108–21083. https://doi.org/10.1109/CVPRW.2018.00272

Segments.ai. (2023). *Segments.ai*. https://segments.ai/point-cloud-labeling.

Song, H., Jo, K., Cho, J., Son, Y., Kim, C., & Han, K. (2022). A training dataset for semantic segmentation of urban point cloud map for intelligent vehicles. *ISPRS Journal of Photogrammetry and Remote Sensing, 187*, 159–170. https://doi.org/10.1016/j.isprsjprs.2022.02.007.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., … Li, J. (2020). Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020*, 797–806.

Tchapmi, L. P., Choy, C. B., Armeni, I., Gwak, J., & Savarese, S. (2017). *SEGCloud: Semantic Segmentation of 3D Point Clouds*. arXiv. https://doi.org/10.48550/ARXIV.1710.07563.

Thomas, H., Qi, C., Deschaud, J.-E., Marcotegui, B., Goulette, F., & Guibas, L. (2019). *KPConv: Flexible and Deformable Convolution for Point Clouds*. 6410–6419. https://doi.org/10.1109/ICCV.2019.00651.

Uchida, T., Hasegawa, K., Li, L., Adachi, M., Yamaguchi, H., Thufail, F. I., … Tanaka, S. (2020). Noise-robust transparent visualization of large-scale point clouds acquired by laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing, 161*, 124–134. https://doi.org/10.1016/j.isprsjprs.2020.01.004.

Vallet, B., Brédif, M., Serna, A., Marcotegui, B., & Paparoditis, N. (2015). TerraMobilita/iQmulus urban point cloud analysis benchmark. *Computers & Graphics, 49*, 126–133. https://doi.org/10.1016/j.cag.2015.03.004.

Virtanen, J.-P., Daniel, S., Turppa, T., Zhu, L., Julin, A., Hyyppä, H., & Hyyppä, J. (2020). Interactive dense point clouds in a game engine. *ISPRS Journal of Photogrammetry and Remote Sensing, 163*, 375–389. https://doi.org/10.1016/j.isprsjprs.2020.03.007.

Xie, Y., Tian, J., & Zhu, X. X. (2020). Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geoscience and Remote Sensing Magazine, 8*(4), 38–59. https://doi.org/10.1109/MGRS.2019.2937630

Xu, Y., Tong, X., & Stilla, U. (2021). Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry. *Automation in Construction, 126*, Article 103675. https://doi.org/10.1016/j.autcon.2021.103675.

Yan, X., Zheng, C., Li, Z., Wang, S., & Cui, S. (2020). PointASNL: Robust Point Clouds Processing using Nonlocal Neural Networks with Adaptive Sampling. *CoRR*, *abs/2003.0*. https://arxiv.org/abs/2003.00492.

Zhang, J., Li, X., Zhao, X., & Zhang, Z. (2022). LLGF-Net: Learning Local and Global Feature Fusion for 3D Point Cloud Semantic Segmentation. In *Electronics* (Vol. 11, Issue 14). https://doi.org/10.3390/electronics11142191.

Zhang, J., Zhao, X., Chen, Z., & Lu, Z. (2019). A Review of Deep Learning-Based Semantic Segmentation for Point Cloud. *IEEE Access, 7*, 179118–179133. https://doi.org/10.1109/ACCESS.2019.2958671

Zhang, Z., Hua, B.-S., & Yeung, S.-K. (2019). ShellNet: Efficient Point Cloud Convolutional Neural Networks using Concentric Shells Statistics. *CoRR*, *abs/1908.0*. http://arxiv.org/abs/1908.06295.

Zhao, H., Jiang, L., Fu, C.-W., & Jia, J. (2019). PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5565–5573.

Zhu, J., Gehrung, J., Huang, R., Borgmann, B., Sun, Z., Hoegner, L., … Stilla, U. (2020). TUM-MLS-2016: An Annotated Mobile LiDAR Dataset of the TUM City Campus for Semantic Point Cloud Interpretation in Urban Areas. *Remote Sensing, 12*(11). https://doi.org/10.3390/rs12111875