A REPRODUCIBLE APPROACH TO ESTIMATE INDOOR SPACE AREA ON A HANDHELD LIDAR DATASET USING DBSCAN

B. Anbaroğlu [1], H. B. Karabay [1]

[1] Geomatics Engineering Department, Hacettepe University, Ankara, Turkey – (banbar, hkarabay)@hacettepe.edu.tr

KEY WORDS: mobile LiDAR, computational reproducibility, digital twin, benchmark dataset, area estimation.

ABSTRACT:

Research on handheld LiDAR data has recently been proliferated due to the emergence of digital twins and indoor mapping. However, most of the existing studies cannot be reproduced in another computational environment. Computational reproducibility requires data, code/software, and computational environment (e.g. versions, settings, etc.) to be openly available. Although there are an increasing number of researches that contribute towards open data, there are still few studies investigating the remaining two aspects. One of the common tasks in digital twin research is the estimation of indoor space areas. This paper contributes to the computational reproducibility of estimating the area of indoor spaces on a handheld LiDAR dataset using the DBSCAN algorithm. The collected dataset -representing the Geomatics Engineering Department of Hacettepe University, code, and the computational environment was made openly available to satisfy the requirements of computational reproducibility. Three different experiments have been carried out: i) identification of the optimal DBSCAN parameter values for a single indoor space, ii) evaluating to what extent these values are applicable to other rooms, and iii) investigating the effect of room enter/exit times on the estimated room sizes. The main finding of this paper is that the simple consideration of an *open-door*, which reduces data collection time, the uncertainty of a wall's coordinates, and imperfect choice of DBSCAN parameters, may substantially increase the estimated indoor space size ranging between approximately 40% to 300%. Consequently, relying solely on the DBSCAN algorithm for indoor space area estimation should not be considered as a valid approach.

Code and computational environment: https://github.com/banbar/HU_Geomatics_LiDAR
Data: https://doi.org/10.6084/m9.figshare.24866175.v1

1. INTRODUCTION

Open science is crucial to the advancement of knowledge, innovation, and the collective well-being of society. Only by relying on an open-science approach, computational reproducibility of research findings could be satisfied (McKiernan et al., 2016). In order to contribute towards open-science, and satisfy computational reproducibility three components of a research must be made publicly accessible: i) data, ii) code/software, and the iii) computational environment (e.g. operating system, library/package versions, etc.) that the research has initially been executed (Bajorath, 2023). Although most research on open-science and computational reproducibility has been on life sciences, geographical information science is increasingly recognising the importance of these principles (Kedron, Li, Fotheringham, & Goodchild, 2021).

International organizations such as ISPRS or ACM has already contributed towards this agenda by providing benchmark datasets (Sithole & Vosselman, 2004) or designing GIS Cups that detail a problem by providing the required datasets (Ali, Krumm, Rautman, & Teredesai, 2012). In addition, an increasing number of scientific journals, such as *SoftwareX* or *The Journal of Open Source Software* are dedicated to the advancement of open-science and computational reproducibility as researchers share their scientific software. Furthermore, other renowned journals including the *International Journal of Geographical Information Science* require a 'Data and Codes Availability Statement', in which the authors are invited to share data and codes used in their research.

The proliferation of Light Detection and Ranging (LiDAR) sensors enabled researchers to collect and analyse point cloud data. The applications that rely on point-cloud data range from autonomous vehicles (Caesar et al., 2020) to archaeology (Chase et al., 2011), and from forestry (Hyyppä et al., 2008) to 3D city modelling (Özdemir & Remondino, 2018). Most of the studies relied on aerial, mobile or terrestrial LiDAR. However; handheld LiDAR sensors are getting popular as well, thanks to the advancement of technology that incorporated LiDAR sensors into smartphones (Catharia et al., 2023; Luetzenburg, Kroon, & Bjørk, 2021).

Handheld LiDAR devices have been commonly used in various research areas, and specifically regarding digital twins, where researchers aim to create highly detailed and accurate 3D representations of physical environments, such as buildings and conference rooms. By capturing precise locational data through laser pulses emitted from the handheld LiDAR system, researchers can generate dynamic and real-time digital replicas of physical spaces. This technology proves particularly valuable in infrastructure management, and Building Information Modelling (BIM). Converting a point cloud data into an as-is BIM is referred to as *scan-to-BIM*; which is usually a labour intensive and error-prone task that necessitates manual operations (Xiong, Adan, Akinci, & Huber, 2013). Although most of the existing research focused on segmentation (i.e. identifying walls, ceiling, floor etc.) and reconstruction of building interiors from point cloud data, it is equally important to obtain geometric constructs, on top of these or separately, such as the indoor space size. Such geometric constructs would be valuable to assess how well the constructed indoor space matches with the building plans.

On the other hand, there are relatively few resources that utilise handheld LiDAR for both capturing both outdoor and indoor of a building. The aim of this paper is to bridge this gap, and provide an openly available dataset of the Geomatics Engineering Department of Hacettepe University, which is scanned first from outside, and then inside. In addition, the developed Python code could readily be used to estimate the area of an indoor space by relying on the DBSCAN algorithm (Ester, Kriegel, Sander, & Xu, 1996). The structure of this paper is as follows. Second section describes the methodology of the paper. Specifically, it first introduces the openly available Hacettepe University, Geomatics Engineering Department handheld LiDAR dataset, and then describes the method to identify the indoor space area of a selected room or lecture hall. Third section describes the results. Fourth section is the discussion, which emphasise on the importance of correct selection of the input parameters of the DBSCAN algorithm (namely *minPts* and *eps*). Finally, in the fifth section conclusions and future research directions are stated.

2. METHOD

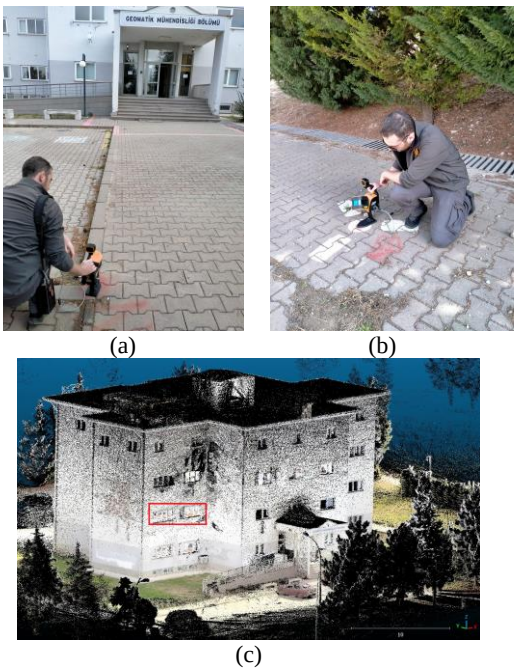
This section describes the method of the paper. First, a new open LiDAR dataset is described. Second, how DBSCAN could and have previously been used for clustering point clouds has been detailed.

2.1 A New and Open Handheld LiDAR Dataset

The Geomatics Engineering Department of Hacettepe University is located at the Beytepe Campus, which is a vibrant campus in Ankara, capital of Turkey. The centre coordinate of the building in latitude and longitude is 39.865566 and 32.733853. The data collection was carried out on 9 October 2022 Sunday with a GeoSLAM ZEB Horizon with its camera ZEB Vision attached. The specifications of the LiDAR scanner, as stated in its web-site, are as follows: i) 300K points/second, ii) a total of 16 sensors, iii) relative accuracy of up to 6mm, and iv) a range of 100 metres (GeoSLAM, 2023). The data collection date was specifically chosen to be a Sunday to ensure a smooth data collection process.

The data collection relied on loop-closure by fixing the sensor at a known point for about 30 seconds (Figure 1a) and then moving around the building to capture its facades. While doing so we have surveyed two more positions (one of them is illustrated in Figure 1b), and then moved inside the building, and completed the loop-closure by finalising the surveying at the initial point. The building has five floors: ground, three above the ground, and the basement. Apart from the basement floor, all of the accessible lecture halls and indoor spaces have been surveyed. In order to ease the data collection process, rooms of the indoor spaces were left open. On the ground floor, three lecture halls and one reception room were surveyed. On the first floor, a lecture hall was surveyed. On the second floor, an office space, PhD meeting room, and a lab was surveyed. Finally, the kitchen was surveyed on the third floor. The two restrooms and a utility room were surveyed only on the ground floor and the first floor.

Once the data collection was finalized it has to be processed to obtain the Laz file. This process resulted in a total three folders and 33 files. The two of the folders are relevant as they contain the images. Specifically, the *Project_1* folder includes all the fisheye images obtained from the camera while data collection, and the *panoromas* folder includes the panoramic images obtained by processing the fisheye images. The image timings of the panoramic images are recorded under the *imageTimings.json* file. This file would be used within the context of this research to identify the start/end times of the scan of an indoor space. The Laz files could be visualised and processed with CloudCompare, which is an open-source software to analyse point clouds (Figure 1c). The lecture hall surveyed in the first floor is highlighted in a red rectangle, as its area was estimated in section 3. The final important file used is the *data.txt*, is obtained by exporting the Laz file in txt format in CloudCompare. The file is composed of eight attributes to define each point (a total of ~131M points). Specifically, the *x*, *y* and *z* of the point define the location of the point, as well as the associated colour information in *r*ed, *g*reen and *b*lue bands are obtained through associating point cloud with the images. Finally, the *t*iming and *i*ntensity of the measurement are also recorded.



(a) (b)

(c)

Figure 1. Handheld LiDAR data collection requires fixation at known points (a, b). Once the collected data are processed, the resulting .laz file can visualised in CloudCompare (the classroom 4 -C4- is highlighted in red) (c).

2.2 DBSCAN for Determining an Indoor Space

The approach taken to determine the indoor space area is to segment the point cloud belonging to that area by a clustering algorithm. One of the common clustering algorithms that has been widely used is the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996). DBSCAN groups data points based on their density within the feature space. The algorithm identifies clusters by relying on two parameters: *eps* and *minPts*, and classifying each data point into three classes: i) core point, ii) border point, and iii) noise point. The algorithm works by iteratively expanding clusters around core points, which have a minimum number of neighbouring data points (*minPts*) within a specified radius (*eps*). Border points are within the *eps* radius of a core point, which are considered part of the same cluster. On the other hand, points that fall outside the radius of all core points are classified as noise points.

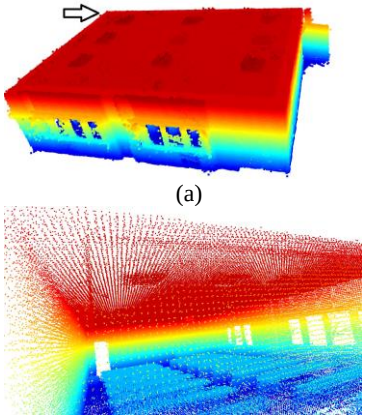
DBSCAN can handle outliers and capable of discovering clusters with varying shapes and sizes. It does not require an input specifying the predefined number of clusters, which makes it suitable for applications in spatial data analysis and anomaly detection. Even though there has been criticism on the algorithm regarding its computational complexity (Gan & Tao, 2015), these issues have been clarified later on (Schubert, Sander, Ester, Kriegel, & Xu, 2017). DBSCAN is considered to be one of the foundational algorithms in which researchers developed various extensions to overcome its limitations, such as handling varying densities within a dataset (Khan, Rehman, Aziz, Fong, & Sarasvady, 2014). The algorithm has uses not only in vector spatial data, but also raster images (Shen et al., 2016), and recently point cloud data (Tao et al., 2015).

Ghosh & Lohani (2013) compared two renowned clustering algorithms -namely DBSCAN and OPTICS- on LiDAR data. They found out that the DBSCAN performed better on the Adjusted Rand Index (Hubert & Arabie, 1985). An analytical study was suggested to adjust the *eps* parameter, which was varied between 0.7 and 4.0 metres on an aerial LiDAR dataset having an average point density of 4.85 points/m^2. A mobile LiDAR dataset was collected on a road had an average point density of approximately 1020 points/m^2. However, how this estimation was realised has not been described in the paper. The average point density could be estimated for an aerial LiDAR survey by dividing the total number of points to the scanned area. On the other hand, it is difficult to have an average point density on a point cloud data obtained through a handheld LiDAR sensor due to the substantial heterogeneity within the data collection procedure. Specifically, the variations in indoor space areas, walking time through these spaces, and the distance between the operator and the concrete objects (e.g. walls and ceiling) as well as the distribution of objects (e.g. desks, furniture, vehicles for parking areas) in space make it difficult to come up with a single average point density measure (Romero-Jarén & Arranz, 2021). Furthermore, in our case, the building was scanned from both outdoor and indoor, which associates points belonging to walls from both outdoor and indoor environments.

The user has to explore the *panoromas* folder and identify when the operator got into and out of the indoor space to be analysed. The user would then note these image numbers, and then the timings would automatically be retrieved from the *imageTimings.json* file. This would enable the extraction of points that are obtained within that time interval. This would be used as an input to the DBSCAN algorithm, which is used to find the most dominant cluster representing the extents of the chosen indoor space (Figure 2).

Figure 2. DBSCAN is used to remove noise from the initial point cloud data

Finally, the point cloud representing the most dominant cluster are excluded for further analysis (Figure 3a). The most dominant cluster is assumed to represent the indoor space to be examined. However, if the DBSCAN parameters are wrongly provided, all points may be considered as noise, in which case the estimated area would be extremely large.



(a)

(b)

Figure 3. The cluster having the highest number of points is assumed to represent the indoor space (a), in which many details including desks, chairs and windows are pertained (b)

Although previous studies suggest an automatic way of estimating the *eps* value, the user would still need to identify the *minPts* (Wang et al., 2019).

3. RESULTS

This section describes the results obtained by analysing the point cloud data identify the area of indoor spaces. The experiments are carried out under four headings. First, the most effective configuration of the DBSCAN parameters has been investigated for a given indoor space. The histogram of the cluster of points representing the investigated indoor space are further examined, which could then be used to have a refined area estimation. Second, an investigation has been carried out to determine the extent to which these values can be generalised. Specifically, six additional indoor spaces are further investigated with the same parameters that have been used in the first experiment. Third, the effect of altering the chosen image time on the results have been investigated for four restrooms that have the same area.

3.1 Adjusting the Parameters of DBSCAN

The two parameters of the DBSCAN algorithm are: *minPts* and *eps*. In order to determine the values of these two parameters an empirical investigation has been carried out for the classroom 4 (i.e. C4 is highlighted in Figure 1c), in which final year students take most of their courses. The *minPts* parameter has been fixed to 60, and the *eps* parameter was varied between 0.15 and 0.25. The actual classroom size of C4 is 71.81 m^2. All of the experiments took almost six minutes on a computer with the following specifications: 16 GB RAM with an Intel Core i7-6500U CPU with 2.50GHz.

There a total number approximately 4.5M points that fall within the C4. Once this point cloud was down sampled with the *voxel_down_sample* method with a voxel size of 0.05 (which has been fixated for all the following experiments), there were 258952 points. The higher the voxel size, the lower the total number of points. These points were then clustered using the *cluster_dbscan* method of the *open3d* package (StackOverflow,

2023). Once all the clusters are formed, the one with the largest number of points was assumed to represent the cluster. The estimated area, number of points within the main cluster (i.e. the cluster having the highest number of points), and the number of clusters are illustrated in Table 1.

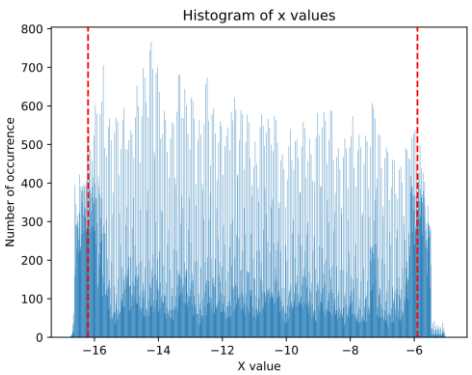
The lowest *eps* value resulted in an unusually large estimated indoor space area. This is because noisy observations were also considered within the cluster, and therefore the area was estimated based on these outliers. In addition, the highest number of clusters also occurred within this context.

Table 1. The effect of the *eps* parameter on the results

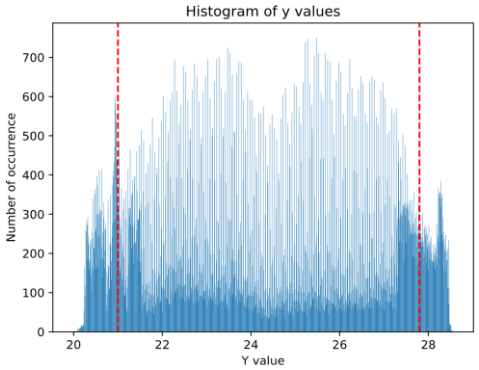
eps	Estimated m^2	# cluster points	# clusters
0.15	5447.01	103,008	275
0.16	100.17	178,865	45
0.17	100.17	199,258	28
0.18	100.61	213,642	28
0.19	101.77	224,582	15
0.20	**101.90**	**228,848**	**10**
0.21	102.94	229,632	5
0.22	102.94	229,788	8
0.23	102.94	229,933	8
0.24	103.32	230,070	11
0.25	104.02	230,114	10

It should also be noted that as expected the number of clusters reduce with incrementing *eps* values, as the formed clusters had to be denser. However, when the value was increased from 0.21 to 0.22, the number of clusters also increased. The reason for this outcome is that noise points might have found sufficient neighbours to form a cluster. Nevertheless, in order to keep the estimated indoor space area at minimum and number of clusters together in balance, the *eps* value was chosen to be 0.20. Consequently, the estimated area was over the real value (i.e. 71.81 m^2) by about 40%. In order to understand this large deviation, the point-cloud corresponding to the main cluster that represents the investigated classroom is further investigated.

The histogram values of the point cloud in all dimensions reveal important insights into the detected cluster. The histogram plots would be saved under the directory that has the dataset and the code, if the *visualize* and *writeout* parameters are enabled in the configuration file. In the case of C4, when *eps* = 0.20 and *minPts* = 60, the histograms along the *x*- and *y*-dimensions are illustrated in Figure 4.



(a)



(b)

Figure 4. The histograms of the x- (a), y- (b) and z-values of the main cluster representing the selected indoor space are also obtained to enable further manual interpretation. The red-lines indicate the selected locations for the wall surfaces.

Once the walls width is eliminated with a manual interpretation, which are visualised in red lines, the estimated area would be much closer to the real value (i.e. $\Delta x \times \Delta y \sim 10.4 \times 6.8 = 70.72$). It could be observed that there is a dense region towards both ends of the histograms that correspond to the walls, which are almost a metre in length.

The main cluster was also saved as *cleaned.pcd*. Once it is observed, the door was also considered to be part of the indoor space. Since the door-opening is along the *x*-dimension, the door's *x*-values could be observed in Figure 4a, at the very right of the histogram, where the door is actually located (refer to Figure 3a). As it is expected, the number of occurrences for the corresponding *x*-values (i.e. $x \sim -5.5$) are substantially lower than most of the remaining values.

3.2 Same Parameters, Different Rooms

This subsection describes how well the previous parameter setting (i.e. *eps* = 0.20, *minPts* = 60 and *voxel size* = 0.05) could be generalised for the different indoor spaces. Specifically, the results regarding seven different indoor spaces are investigated in this subsection. The investigated room IDs; C1, C2, and C3 correspond to the classrooms at the first-floor of the building. The C4 is above the C1 but without the extra space for the hangers. BA is the office space of the first author, PhD is the meeting room of graduate students. The BA, PhD, and GIS Lab are on the second-floor, and finally Kitchen is on the third-floor. The ratio column identifies to what extent the main cluster used to estimate the room size is within the sampled point-cloud. For example, in the C4 case described in Table 1, this ratio value would have been 228848 / 258952 =0.88. The results are illustrated in Table 2.

Table 2. The estimation of indoor space areas on different rooms

Room ID	True m^2	Estimated m^2	# clusters	Ratio
C1	77.05	106.60	11	0.85
C2	66.31	95.12	34	0.69
C3	66.31	117.83	19	0.78
BA	21.7	42.48	22	0.79
PhD	29.78	63.63	16	0.87
Kitchen	14.77	23.69	12	0.74
GIS Lab	49.33	76.25	5	0.94

The running time of these experiments varied between three minutes and eight minutes. This run time cannot be directly associated with the total cluster points, true room size, or the total number of down sampled points. All of these, in addition to the distribution of sampled points within the room contribute towards the run-time in a complex way. Similar to the previous scenario, the area is overestimated with a percentage ranging from 38% (i.e. room ID: C1) to almost 215% (i.e. room ID: PhD). As the DBSCAN parameters used in this experiment were derived from C4, directly above C1, they provided the closest approximation to C1. On the other hand, the error in the PhD case is substantial, and suggest that the one parameter setting cannot be directly applied to all rooms.

3.3 The Effect of Marginal Changes on Room Enter/Exit Times

The rooms have previously been identified by user selecting the appropriate images from the *panoromas* folder. However, different users may select different images, and the effects of this decision on area estimation are investigated on this section. Specifically, four restrooms (two on the ground-floor, and two on the first-floor) are analysed in detail in this subsection. While male and female restrooms are identical within themselves (i.e. M0 ≡M1 and F0 ≡F1, where the M and F refer to male and female respectively, while the numbers refer to the floor of the restroom). Similar to the previous scenario, the DBSCAN parameters were set as *eps* = 0.20 and *minPts* = 60, with a down sample voxel size of 0.05.

The optimal start and end times are indicated with two parameters: *s* and *e*, which vary depending on the specific restroom. This optimal decision was altered with one image on both aspects. The optimal selection of a room is denoted with *s* and *e* denoting start and end respectively. Therefore, *s*-1 and *e*+1 scenario would indicate the user has selected the start image one prior compared to our optimal selection, and the user selected the exit image one after with respect to our optimal selection. The estimated areas are illustrated for these different scenarios in Table 3. All of the surveyed restrooms have the same area (i.e. 9.18 m^2) on the building plan.

Table 3. Estimated area for the surveyed restrooms

Scenario		M0	F0	M1	F1
s-1	e-1	13.83	19.13	15.15	21.1
s	e-1	13.86	18.47	15.1	14.11
s+1	e-1	13.82	17.21	15.08	14.08
s-1	e	13.83	19.12	15.48	21.1
s	**e**	**13.86**	**18.62**	**15.39**	**14.11**
s+1	e	13.8	18.42	15.49	14.07
s-1	e+1	14.12	27.2	15.48	21.1
s	e+1	14.11	19.14	15.43	19.68
s+1	e+1	14.11	18.87	15.48	14.07

The results are consistent at different scales. The male restrooms are quite stable regardless of the scenario, while the one on the ground floor produced more accurate results. Similar to the previous analysis described in subsection 3.2, area is over-estimated with about 50% for the male restroom at the ground floor, and with about 65% for the male restroom at the first floor. On the other hand, a lesser level of consistency was observed for the female restrooms. Consequently, the over-estimation was even more emphasised in this context, and the in worst-case reached to almost 300% for the female restroom on the ground floor.

4. DISCUSSION

The estimated areas using the manually tuned DSBSCAN parameters are substantially larger than the true values. The main reasons for this outcome are two-fold. First, the area was estimated on the simple calculation of the *width* × *depth* of the room along the *x*- and *y*- dimensions. These values are assumed to be their maximum, which meant the inclusion of all points representing the wall into the estimated area. Since all the rooms are also scanned from outside of the building, this region could be as large as a metre, which could be observed in Figure 4. Second, the door of a room has also been considered to be part of the indoor space, which also increased the estimated area. The one-size-fits-all approach, in which tuning the parameters of DBSCAN for a single indoor space, and then using them in the remaining rooms was found to be an invalid approach. Specifically, tuning the DBSCAN for different rooms is required. This finding is in-line with (Lari & Habib, 2012), who suggested the estimation of local point density indices. Specifically, different rooms may have different densities within themselves, which may be required to be considered.

Planar surface detection is one of the potential research areas that could contribute to this research, in addition to investigating different clustering algorithms such as Fuzzy C-means (Biosca & Lerma, 2008). Furthermore, recent research investigated how Haugh Transform could be used for the detection of planar surfaces in a point cloud dataset (Tian et al., 2020). The HT idea could be extended to include other classes in a more complex outdoor scene including trees and pedestrians by leveraging convolutional neural networks (Song et al., 2020). One of the challenging tasks on planar surface detection on a point-cloud dataset is the identification of parallel planes. In order to avoid this problem, Walczak, Poreda, & Wojciechowski (2019) utilised an improved version of DBSCAN (namely HDBSCAN). They have relied on the S3DIS benchmark dataset, which has been kindly provided by researchers at Stanford University (Armeni et al., 2016). The seminal work provides a framework for semantic parsing of the point cloud, which could segment objects like a board, bookcase or table, within a room. Furthermore, Nguyen, Belton, & Helmholz (2019) investigated the effectiveness of other benchmark models for planar surface detection including Random Sample Consensus (RANSAC), Principal Component Analysis (PCA) or Robust and Diagnostic Principal Components Analysis (RDPCA).

5. CONCLUSIONS

This paper contributed towards the estimation of an indoor area space on a point cloud dataset obtained from a handheld LiDAR sensor using the DBSCAN algorithm. The findings of this paper are reproducible, as both the data and the code has been made openly available. The approach taken in this paper could be considered as a base-line method, as new methods would likely improve the accuracy of the estimated indoor space areas. Planar surface detection should be incorporated into the analysis, and the developed reproducible approach should be generalised into a Python package by incorporating existing methods. Further tests should be carried out both on this dataset and on existing openly available datasets, such as S3DIS to investigate the effectiveness of the proposed methods.

ACKNOWLEDGEMENTS

We would like to thank Berk Alp Direm who has facilitated the data collection process as well as data pre-processing.

REFERENCES

Ali, M., Krumm, J., Rautman, T., & Teredesai, A. (2012). ACM SIGSPATIAL GIS Cup 2012. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, 597–600. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2424321.2424426

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

Bajorath, J. (2023). Data and code availability requirements in open science and consequences for different research environments. *Artificial Intelligence in the Life Sciences*, *4*, 100085. https://doi.org/10.1016/j.ailsci.2023.100085

Biosca, J. M., & Lerma, J. L. (2008). Unsupervised robust planar segmentation of terrestrial laser scanner point clouds based on fuzzy clustering methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*(1), 84–98. https://doi.org/10.1016/j.isprsjprs.2007.07.010

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., … Beijbom, O. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628. https://doi.org/10.1109/CVPR42600.2020.01164

Catharia, O., Richard, F., Vignoles, H., Véron, P., Aoussat, A., & Segonds, F. (2023). Smartphone LiDAR Data: A Case Study for Numerisation of Indoor Buildings in Railway Stations. *Sensors*, *23*(4), 1967. https://doi.org/10.3390/s23041967

Chase, A. F., Chase, D. Z., Weishampel, J. F., Drake, J. B., Shrestha, R. L., Slatton, K. C., … Carter, W. E. (2011). Airborne LiDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. *Journal of Archaeological Science*, *38*(2), 387–398. https://doi.org/10.1016/j.jas.2010.09.018

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231. Portland, Oregon: AAAI Press.

Gan, J., & Tao, Y. (2015). DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 519–530. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2723372.2737792

GeoSLAM. (2023). ZEB Horizon: The Ultimate Mobile Mapping Solution. Retrieved 1 October 2023, from GeoSLAM website: https://geoslam.com/wp-content/uploads/2021/02/ZEB-Horizon-User-Manual-v1.3.pdf

Ghosh, S., & Lohani, B. (2013). Mining lidar data with spatial clustering algorithms. *International Journal of Remote Sensing*, *34*(14), 5119–5135. https://doi.org/10.1080/01431161.2013.787499

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., & Maltamo, M. (2008). Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing*, *29*(5), 1339–1366. https://doi.org/10.1080/01431160701736489

Kedron, P., Li, W., Fotheringham, S., & Goodchild, M. (2021). Reproducibility and replicability: Opportunities and challenges for geospatial research. *International Journal of Geographical Information Science*, *35*(3), 427–445. https://doi.org/10.1080/13658816.2020.1802032

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 232–238. https://doi.org/10.1109/ICADIWT.2014.6814687

Lari, Z., & Habib, A. (2012). Alternative Methodologies for the Estimation of Local Point Density Index: Moving Towards Adaptive Lidar Data Processing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXIX-B3*, 127–132. https://doi.org/10.5194/isprsarchives-XXXIX-B3-127-2012

Luetzenburg, G., Kroon, A., & Bjørk, A. A. (2021). Evaluation of the Apple iPhone 12 Pro LiDAR for an Application in Geosciences. *Scientific Reports*, *11*(1), 22221. https://doi.org/10.1038/s41598-021-01763-9

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., … Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*, e16800. https://doi.org/10.7554/eLife.16800

Nguyen, H. L., Belton, D., & Helmholz, P. (2019). Planar surface detection for sparse and heterogeneous mobile laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *151*, 141–161. https://doi.org/10.1016/j.isprsjprs.2019.03.006

Özdemir, E., & Remondino, F. (2018). Segmentation of 3D Photogrammetric Point Cloud for 3D Building Modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4-W10*, 135–142. https://doi.org/10.5194/isprs-archives-XLII-4-W10-135-2018

Romero-Jarén, R., & Arranz, J. J. (2021). Automatic segmentation and classification of BIM elements from point clouds. *Automation in Construction*, *124*, 103576. https://doi.org/10.1016/j.autcon.2021.103576

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*, *42*(3), 19:1-19:21. https://doi.org/10.1145/3068335

Shen, J., Hao, X., Liang, Z., Liu, Y., Wang, W., & Shao, L. (2016). Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm. *IEEE Transactions on Image Processing*, *25*(12), 5933–5942. https://doi.org/10.1109/TIP.2016.2616302

Sithole, G., & Vosselman, G. (2004). Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *59*(1), 85–101. https://doi.org/10.1016/j.isprsjprs.2004.05.004

Song, W., Zhang, L., Tian, Y., Fong, S., Liu, J., & Gozho, A. (2020). CNN-based 3D object classification using Hough space of LiDAR point clouds. *Human-Centric Computing and Information Sciences*, *10*(1), 19. https://doi.org/10.1186/s13673-020-00228-8

StackOverflow. (2023, August 25). Answer to 'open3d voxel downsampling then dbscan cluster then approximately upscale clustering to the original image'. Retrieved 21 December 2023, from https://stackoverflow.com/a/76975970/1959766

Tao, S., Wu, F., Guo, Q., Wang, Y., Li, W., Xue, B., … Fang, J. (2015). Segmenting tree crowns from terrestrial and mobile LiDAR data by exploring ecological theories. *ISPRS Journal of Photogrammetry and Remote Sensing*, *110*, 66–76. https://doi.org/10.1016/j.isprsjprs.2015.10.007

Tian, Y., Song, W., Chen, L., Sung, Y., Kwak, J., & Sun, S. (2020). Fast Planar Detection System Using a GPU-Based 3D Hough Transform for LiDAR Point Clouds. *Applied Sciences*, *10*(5), 1744. https://doi.org/10.3390/app10051744

Walczak, J., Poreda, T., & Wojciechowski, A. (2019). Effective Planar Cluster Detection in Point Clouds Using Histogram-Driven Kd-Like Partition and Shifted Mahalanobis Distance Based Regression. *Remote Sensing*, *11*(21), 2465. https://doi.org/10.3390/rs11212465

Wang, C., Ji, M., Wang, J., Wen, W., Li, T., & Sun, Y. (2019). An Improved DBSCAN Method for LiDAR Data Segmentation with Automatic Eps Estimation. *Sensors*, *19*(1), 172. https://doi.org/10.3390/s19010172

Xiong, X., Adan, A., Akinci, B., & Huber, D. (2013). Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction*, *31*, 325–337. https://doi.org/10.1016/j.autcon.2012.10.006