A Big Data approach for building basement detection using airborne LiDAR data

Conference Paper · August 2023

CITATIONS
0

READS
51

5 authors, including:

Anh-Vu Vo
University College Dublin
27 PUBLICATIONS **888** CITATIONS
SEE PROFILE

Michela Bertolotto
University College Dublin
266 PUBLICATIONS **4,036** CITATIONS
SEE PROFILE

Debra F Laefer
New York University
223 PUBLICATIONS **4,233** CITATIONS
SEE PROFILE

Nhien-An Le-Khac
University College Dublin
338 PUBLICATIONS **5,515** CITATIONS
SEE PROFILE

ISMLG 2023
4th International Symposium on Machine Learning and Big Data in Geoscience
University College Cork, 29th August – 1st September 2023

A Big Data approach for building basement detection using airborne LiDAR data

Anh Vu Vo[*1], Michela Bertolotto[1], Debra F. Laefer[2], Nhien-An Le-Khac[1] and Ulrich Ofterdinger[3]

[1]School of Computer Science, University College Dublin, Ireland

[2]Center for Urban Science and Progress, New York University, United States

[3]School of Natural and Built Environment, Queen's University Belfast, Northern Ireland

[*]presenting author (email: anhvu.vo@ucd.ie)

Keywords: basement, detection, lidar, distributed computing

1 INTRODUCTION

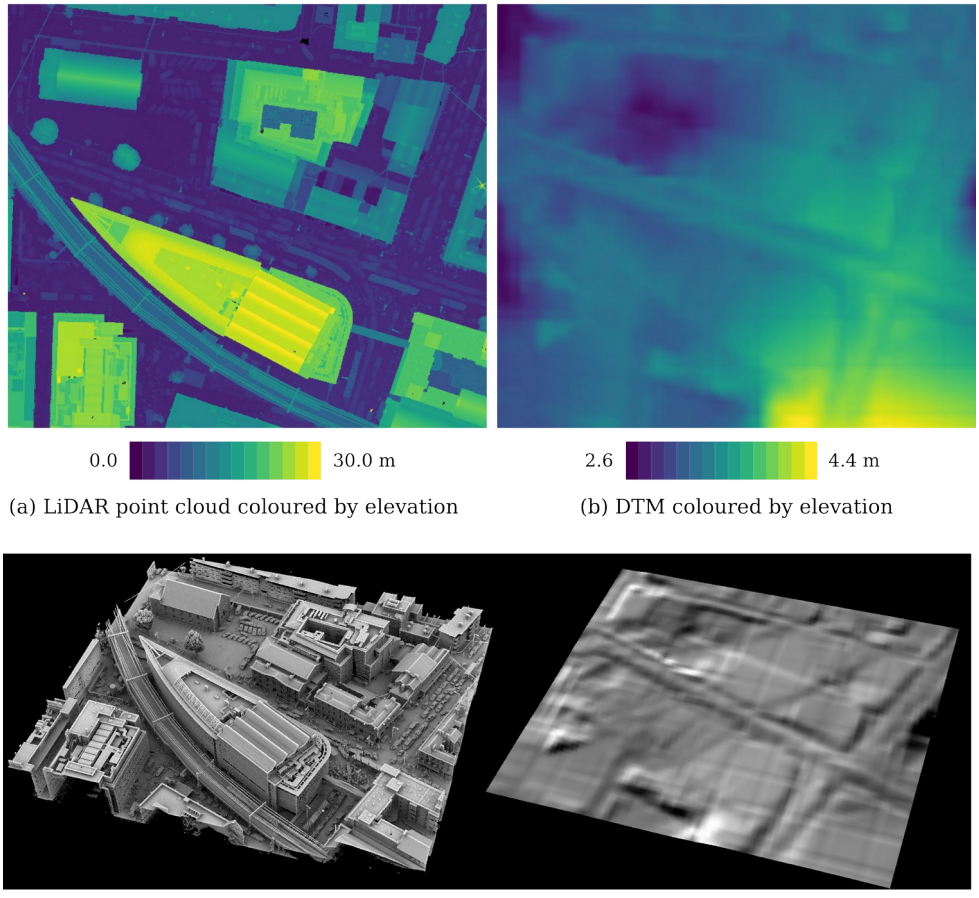
Information about building basements is important to local authorities for applications such as planning, risk assessment, and taxation [1]. The 2005 Greater Dublin Strategic Drainage Study[1] highlighted both the lack of basement information in the Dublin region in Ireland and the criticality of having such information for flood risk management. This paper presents an approach for automatically extracting basement information from remote sensing data to enhance the current knowledge of basement structures in Dublin. Specifically, a large airborne LiDAR (Light Detection and Ranging) dataset of over 1.4 billion points acquired by Laefer et al. in 2015 [2] was integrated with Digital Terrain Model (DTM) data provided by the Office of Public Works (OPW) to detect open basements and identify basement levels. A Big Data algorithm was derived to handle the intensive integration operation which involves two large spatial datasets of different types.

2 METHODOLOGY

Figure 1 shows the two input data sources necessary for the algorithm: a LiDAR point cloud, **P,** (Figures 1a&c) which is a collection of 3D points representing the geometry of the urban environment; and a DTM, **D**, (Figures 1b&d) containing an array of terrain elevation values. The particular point cloud used in this analysis has a point density of over 300 points/m^2. The OPW's DTM was provided at a ground sampling distance of 2.0 m. Essentially, **P** and **D** represent two elevation fields. While **P** contains elevation samples from all above- and under-ground objects, **D** contains only ground points. By comparing elevation samples in **P** with corresponding terrain elevation values in **D**, one can straightforwardly identify LiDAR points lying below the ground level. Notably, point clouds and DTMs are often very large and can

[1]https://www.dublincity.ie/sites/default/files/media/file-uploads/2018-07/GDSDS_Policy_Technical_Document_for_Basements.pdf

contain billions to hundreds of billions of sampling points. Integrating such large datasets can be very time consuming. The algorithm presented in this paper addresses the data intensiveness of the required spatial integration by partitioning and decoupling the input datasets so that different data portions can be analysed in parallel by different computing nodes (i.e. autonomous computers) to reduce the computational time. The computing strategy is often known as data parallelisation. Apache Spark[2], the parallel computing framework widely used for Big Data analytics, was selected for the implementation of the algorithm.



(a) LiDAR point cloud coloured by elevation

(b) DTM coloured by elevation

(c) LiDAR point cloud coloured by ambient occlusion

(d) Hillshaded DTM

Figure 1: Point cloud and Digital Terrain Model

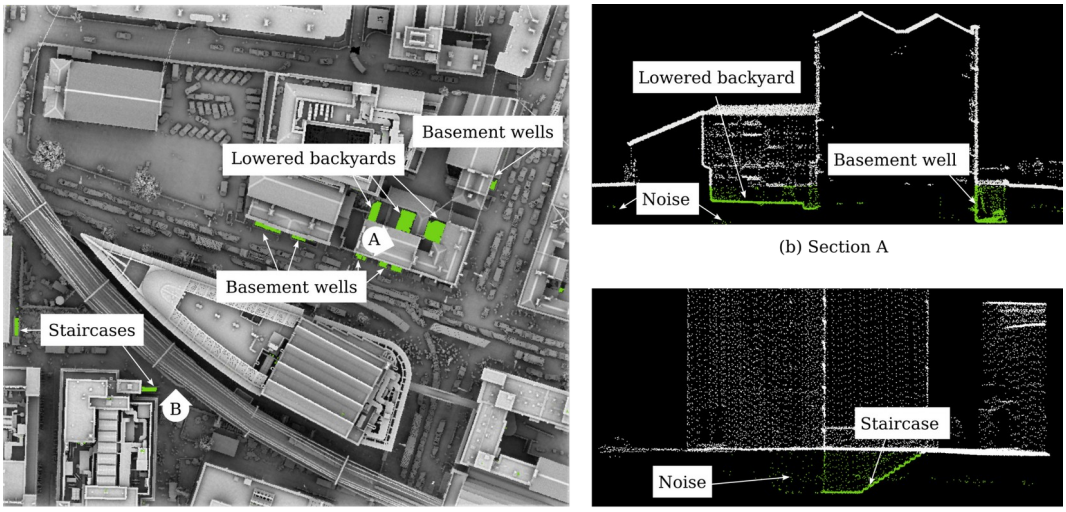
The first step of the algorithm is to define a 2D regular raster grid **G** (e.g. 0.5 x 0.5 m) overlaying the spatial extent of interest. Subsequently, each point $p \in \mathbf{P}$ is mapped to cell $c \in \mathbf{G}$ that contains p and the resulting data are grouped by the raster cell. A similar operation is

[2] https://spark.apache.org/

carried out for each elevation value $v \in \mathbf{D}$. Those operations are formulated so that the data can be distributed using the parallel *map* and *groupBy* functions in Apache Spark. A join operation (i.e. *join* function in Apache Spark) is followed to connect the LiDAR point group in each raster cell with the corresponding terrain elevation values. At this stage, the data subsets in different cells are analysed in parallel to compare the LiDAR points against the median terrain elevation in the cell to identify LiDAR points lying below the terrain level. Subterranean points in the proximity of a building are considered potential basement points from that building. To reduce false detections caused by noisy LiDAR points which are often characterised by a low density, only points having a local density index [3] above a certain threshold (e.g. 20 points/m^2) are kept in the result set. The formulation of the algorithm helps it take full advantage of the *map*, *reduce*, *groupBy*, and *join* functions in Apache Spark to parallelise the algorithm thereby accelerating the computation.

3 RESULTS & DISCUSSIONS

The algorithm described in Section 2 was successfully implemented. Figure 2 presents an example of the results. The green points represent basement points subterranean extracted by the algorithm. As shown in Figure 2a, many features indicative of basement structures such as basement wells, basement staircases, and lowered backyards were correctly detected by the algorithm. Figures 2b&c present closeup views of selected features. The low density noise points, which were later filtered out using the density threshold, are also visible in the close-ups.



(a) Basement detection results

(b) Section A

(c) Section B

Figure 2: Point cloud and Digital Terrain Model

Detailed evaluation was performed for a small group of buildings on Pearse Street using the basement database provided by the Dublin City Council (DCC). Among the twelve

buildings investigated, the algorithm was able to correctly identify four buildings that have basements and five buildings not having a basement. In addition, the algorithm detected basement features (i.e., basement wells and staircases to below-ground levels) in three buildings not documented in the DCC's basement database. Notably, the algorithm provided basement level information, which is valuable information not available in the DCC database.

Given 16 computing nodes, each of which has 8 CPU cores, the integration of 1.4 billion LiDAR points and the corresponding DTM was accomplished in under five minutes. Without parallelism, the computation would have required several hours. That massive reduction in computational time demonstrated the ability of the algorithm to handle big data.

4 CONCLUSIONS

This paper introduces a Big Data approach for automatically extracting basement information (i.e. presence of basements and levels) from LiDAR and DTM data. The proposed approach is fast, scalable, and can handle very large amounts of data due to the use of parallel computing. Experimental results showed that the algorithm could correctly detect many known basement structures and identify features potentially indicative of basement structures not currently documented. One limitation of the presented approach is that only open basements (i.e. those visible from the sky view) can be detected.

ACKOWLEDGEMENTS

Funding for this project was provided by the National Science Foundation as part of the project "UrbanARK: Assessment, Risk Management, Knowledge for Coastal Flood Risk Management in Urban Areas" NSF Award 1826134, jointly funded with Science Foundation Ireland (SFI - 17/US/3450) and a research grant (USI 137) from the Department for the Economy Northern Ireland under the US-Ireland R&D Partnership Programme. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The aerial LiDAR data of Dublin were acquired with funding from the European Research Council [ERC-2012-StG-307836] and additional funding from Science Foundation Ireland [12/ERC/I2534]. The Digital Elevation Model was provided by The Office of Public Works. The computing cluster used for the testing was provided by NYU High Performance Computing Center. The authors would like to thank the NYU HPC staff and OPW staff for their excellent technical support.

REFERENCES

[1] Lieberman, J., Ryan, A. (2017). OGC underground infrastructure concept study engineering report. *OGC Engineering Report*.

[2] Laefer, D. F., Abuwarda, S., Vo, A. V., Truong-Hong, L., Gharibi, H. (2017). 2015 aerial laser and photogrammetry survey of Dublin city collection record.

[3] Vo, A. V., Lokugam Hewage, C. N., Le Khac, N. A., Bertolotto, M., Laefer, D. (2021). A parallel algorithm for local point density index computation of large point clouds. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 8.