ON VOLUME DATA REDUCTION FOR LIDAR DATASETS

K. Becek [1,2,*], P. Boguslawski [1]

[1] Wroclaw University of Science and Technology, Faculty of Geoengineering, Mining and Geology, Wroclaw, Poland - (kazimierz.becek, pawel.boguslawski)@pwr.edu.pl
[2] Dept. of Geomatics Engineering, Zonguldak Bulent Ecevit University, Zonguldak, Turkey, kazimierz.becek@beun.edu.tr

Commission IV, WG IV/1

KEY WORDS: LiDAR, Redundancy, Q-tree, Big Data Problem, Topography, DTM

ABSTRACT:

This paper discusses a current issue for several experimental science disciplines, which is the Big Data Problem (BDP). This research study focused on light intensity and ranging (LiDAR) datasets, which are collected for modelling spatial features found on the surface of the earth. Currently, LiDAR datasets are known to be extremely redundant for many applications. Using a formula that allows for calculating the variance of the target-induced error (so-called *T-error*) caused by the discretisation and quantisation of a 3D surface as a criterion for the quantitative assessment of the fidelity of a model, the use of a Q-tree-based split of the surface is proposed for cells of various sizes depending on the fidelity requirements. A LiDAR dataset representing a 1 km x 1 km terrain surface tile using approximately 12 x 10^6 points was used during the experiments. The initial LiDAR dataset was used to produce a digital terrain model (DTM) at a 0.5 m x 0.5 m resolution, which was used as a reference model. Subsequently, the initial LiDAR dataset was decimated at various rates, and the resulting DTMs were compared with the reference model. The Q-tree based data structure was utilised to illustrate that the Q-tree approach allows for the production of DTMs at a ‘controlled’ fidelity with a considerable reduction in data volume.

1. INTRODUCTION

The Big Data Problem (BDP) is an unwelcome by-product of the acquisition of spatial data at continuously increasing spatial, spectral radiometric and temporal resolution levels (e.g., Jianping et al. 2009). This trend in spatial data acquisition provides several advantages that allow for an increased level of fidelity of geospatial models. One way to reduce the impact of the BDP on business and science fields is to use the precise amount of data that is necessary for a given task, perhaps by observing a certain ‘safety margin’. This is only possible if a criterion exists to effectively translate the assumed fidelity or accuracy level into a procedure to select a subset of data required for the task.

Technology developments in surveying equipment, including laser scanners (LiDAR), have increased the accuracy of captured data and have reduced the time required for the acquisition of data. This allows for an increased fidelity of reality modelling; however, the volume of data produced by the state-of-the-art equipment significantly contributes to the increasing challenge related to data storage and processing time as well as to the management and dissemination of data, which is known as the BDP.

One way to mitigate the BDP is to reduce the volume of the data used for a given purpose. A ‘smart’ approach to reducing or completely removing the redundancy of geodata is to set a quantitative criterion, which helps identify a subset of the entire dataset that would be sufficient to achieve a given goal or to create a product at an assumed fidelity level.

In this paper, the redundancy issue related to LiDAR datasets is examined. The starting point is the observation that to estimate a slope of a surface of a given non-divisible area (a unit), only three LiDAR points are required; however, a LiDAR dataset could store 10 or even more points to represent one square metre of a surface. Therefore, the redundancy in the considered case is at least 70%.

A simulation experiment was conducted, which was designed to identify differences between: a) Digital Terrain Models (DTM)s, which were derived by decimation at arbitrary rates from the original LiDAR dataset, and b) datasets, which were decimated using a Q-tree 2D space partitioning algorithm. The resulting DTMs were compared with the DTM derived from the full resolution LiDAR dataset.

In the ‘a’ case, the decimation was performed by randomly selecting LiDAR points without considering the geometric properties of the modelled surface. In the ‘b’ case, the fixed-σ criterion (Becek, 2012) was used to perform the Q-tree partitioning of the LiDAR dataset and to derive the final DTM as an array of pixels of various sizes, which were multipliers of the smallest pixel (0.5 m x 0.5 m in this case).

The results of the comparison of the DTMs derived using both methods clearly indicated that the Q-tree approach implemented with the fixed-σ criterion performs much better in terms of the volume of data needed to represent a terrain. An unquestionable attribute of the Q-tree approach is that it can be fully controlled by the user in terms of setting a level of accuracy for the final DTM. One drawback of the Q-tree is that contemporary hardware and software solutions implicitly assume that data, e.g. image data, are stored as an array of pixels of equal dimensions. This simply means that the Q-tree to be shown on a screen must be converted into pixels of the same size.

2. MATERIALS AND METHOD

2.1 Test Area

As a test site, a topographically diverse area was selected in which the elevation varied from 1 m to 80 m above mean sea level (amsl). Both flat and hilly terrain features are represented by the selected sample. Narrow, both natural and anthropogenic, channels were included in the sample. Figure 1 shows a

hillshadow picture of a DTM of the 1 km x 1 km area of interest (AOI) selected for the study.

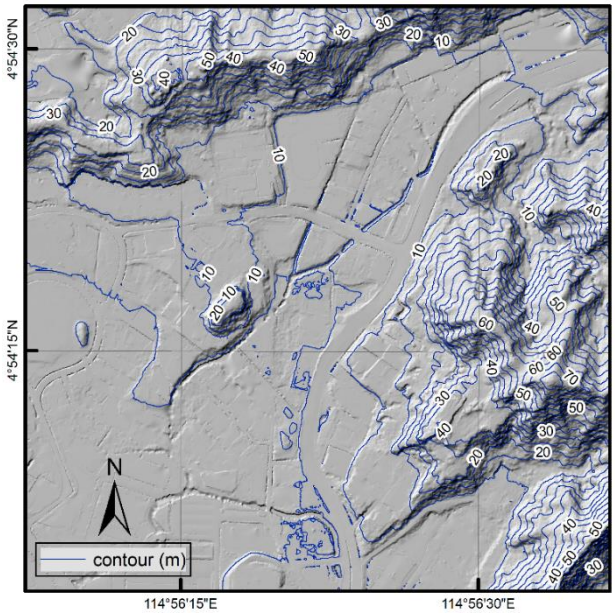


Figure 1. Hillshadow DTM of the AOI

2.2 Data

The primary data used for this experiment were captured during LiDAR and aerial photography acquisition, which took place in Brunei Darussalam on 18 March 2018. The aim of the LiDAR data capture acquisition was to study the features of the urban structure, including vegetation and water bodies. The designated LiDAR point resolution was >10 ptm^{-2}, and orthophoto with pixel size of 5 cm. A RIEGL LMS-Q680i LiDAR Scanner was used for the acquisition. This instrument is a full waveform instrument, allowing for capturing an unlimited number of returns. The instrument is able to scan at 266,000 pts^{-1}, which translates to 200 lines^{-1}. The ranging accuracy of the instrument is 0.02 m (one sigma). The IGI GPS/IMU AeroControl II system was used to provide the navigation and orientation parameters. The acquisition was performed on a DA42 MPP aircraft during good weather conditions from a flying altitude of 600 m above the terrain.

The LiDAR dataset was classified using the TerraSolid software package. The classes of the LiDAR points and some statistics for the identified classes of points are shown in Table 1.

Class	No of pt.	Dens. (ptm^{-2})	Min Z (m)	Med. Z (m)	Max Z (m)
Ground	2,336,142	2.5	0.99	6.43	79.3
Low veg	1,012,185	1.0	1.07	8.14	79.2
Med veg	824,548	0.82	1.42	16.52	81.1
High veg	7,504,178	7.5	3.15	36.08	149.3
Buildings	1,013,509	1	3.56	15.72	87.1
Total	12,690,562	12.7			

Table 1. LiDAR point statistics used during the experiment

The 0.5 m x 0.5 m lattice DTM was produced by dividing the triangulated ground class points only.

2.3 Method

2.3.1 Vertical Accuracy Model for a DTM

A short outline of the vertical accuracy model for a DTM, which was essential to this research, is provided below (Becek, 2008).

The variance of the pixel error of a DTM can be written as the sum of variances of three statistically independent error sources as follows:

$$\sigma_{DEM}^2 = \sigma_I^2 + \sigma_E^2 + \sigma_T^2 \quad (1)$$

where σ_I^2, σ_E^2 *and* σ_T^2 = variance of the instrument - (*I*), environment - (*E*) and target-induced error, respectively.

The *I-error* occurred due to the instrument and method used to capture the data based on which the DTM/DSM in question was captured. The *E-error* occurred due to the environmental conditions/parameters that adversely impact the function of the instrument as well as other equipment during data acquisition. Both components of the *I-* and *E-error* sources usually remain quite stable, especially when the time span of the data capture is relatively short.

The *T-error* occurred due to the geometric properties of the terrain surface and the assumed pixel size of the DTM/DSM. The variance of the *T-error* can be calculated using Equation 2 (Becek, 2008):

$$\sigma_T^2 = \frac{d^2 tg^2(s)}{12} \quad (2)$$

where *d* = pixel size
s = slope.

As terrain topography varies from pixel to pixel, which is reflected in Equation 2 by slope *s*, the *T-error* varies accordingly. The magnitude of the *T-error* is also controlled by pixel size *d*: a larger pixel size enhances the *T-error* variance.

Equation 2 and the conclusions regarding the variable that controlled the magnitude of the *T-error* level formed the basis of the proposed approach to control the redundancy of digital data, and in particular, the redundancy of the LiDAR data representing the surface of the terrain.

2.3.1 Q-tree Data Structure

One of the schemes used to partition a 2D space is known as a quad tree, or a Q-tree (Finkel & Bentley 1974, Samet 1984, de Berg et al. 2008). An algorithm that produces a Q-tree recursively subdivides a 2D space into four quadrants, or regions, also known as leaves. The subdivision of leaves continues until the smallest size of a leaf (assumed) is reached or a space covered by a leaf satisfies a certain criterion. For this project, Equation 2 was used as a criterion, which translated the slope of the terrain within a leaf and the leaf size with the accuracy of the pixel elevation of a DTM (Becek, 2008, 2012, 2014).

2.3.2 The Algorithm

To conduct this study, a slightly different algorithm for generating a Q-tree was used. Rather than dividing the large leaves into four smaller leaves, a reverse, or 'bottom up', procedure was implemented. The major steps and the most important computer implementation details of the algorithm are as follows:

1. The AOI was subdivided into 2^{20} cells (1,048,576), which translates to 0.98 m for one cell on the ground.
2. The elevations of the LiDAR points within each cell were used to estimate the slope.
3. Using Equation 2, the variance of the *T-error* was calculated.
4. The *T-error* variance was compared with the user-defined vertical accuracy level (expressed as the variance of a DTM's vertical error). Five accuracy levels were selected for the simulation experiment.
5. If the *T-error* was smaller than the given vertical accuracy level, four leaves were merged together to form a larger leaf. This step is shown in Figure 2.
6. If the *T-error* estimated for a particular leaf was equal to or larger than the threshold, the leaf was accepted as the final leaf.

The computer implementation of the algorithm was performed using the 64-bit version of Python script.

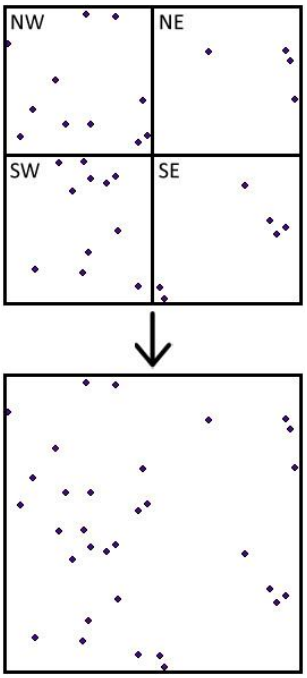


Figure 2. Leaf Merging Operation

To evaluate the validity of the proposed Q-tree partitioning of the LiDAR dataset procedure, five versions of the original LiDAR dataset were produced, and then corresponding DTMs were developed at the same resolution (0.5 m x 0.5 m). Each DTM was compared with the reference DTM (produced from the full resolution of the LiDAR dataset).

Both volume reduction procedures, i.e. the decimation- and Q-tree-based procedures, were compared, and conclusions were drawn.

3. RESULTS

The decimation rate was expressed in terms of the number of LiDAR points omitted from the full resolution dataset. For example, decimation rate = 10 means that every 10^{th} point was taken for the decimated dataset. Because the point density of the ground class was 2.5 ptm^{-2}, the arbitrary selected decimation rates were 10, 50, 100, 200 and 1000. These decimation rates translated to 0.25, 0.05, 0.025, 0.0125 and 0.0025 ptm^{-2}, respectively.

Table 2 lists the basic statistics of the comparison of the decimated DTMs at specific arbitrary selected rates. For all cases, the decimation did not introduce any bias to the decimated models (Mean dZ was close to 0 m) except the doubted reading for the 1000 decimation rate. The standard deviation of the mean dZ increased according to the increased decimation rate, as expected.

No of pixels/ decimation rate	Min dZ (m)	Max dZ (m)	Mean dZ (m)	Overall Std dZ (m)
2336/1000	-38.21	25.68	-0.24	3.21
11680/200	-13.88	11.11	-0.1	1.48
23360/100	-10.57	10.23	-0.06	1.08
46723/50	-8.94	9.82	0.01	0.80
233600/10	-6.33	5.44	0.01	0.36

Table 2. Results of the comparison of the decimated DTMs vs. full resolution DTMs

In the next step, five versions of the Q-tree for the DTMs were derived, assuming the following arbitrary selected accuracy levels (the fixed-σ criterion): 0.36, 0.8, 1.08, 1.48 and 3.21 m. These Q-tree DTMs were compared to the full resolution DTMs. Table 3 lists the basic statistics of the comparisons for each DTM identified according to the number of cells.

Number of cells	Min dZ (m)	Max dZ (m)	Mean dZ (m)	Overall Std dZ (m)
2206	-3.1	3.6	-0.04	0.89
10711	-1.7	1.9	0.0	0.57
17652	-2.5	2.6	0.0	0.49
27380	-1.8	1.7	0.0	0.42
69857	-0.8	0.9	-0.01	0.29

Table 3. Results of the comparisons of Q-tree DTMs vs. full resolution DTMs

Some of the data provided in Tables 2 and 3 were used to produce Figure 3, which shows a relationship between the standard deviation of the mean difference (or the accuracy of the DTM) vs. the number of cells/pixels used to represent the DTM. Corresponding curves for both types of DTM representations are shown.

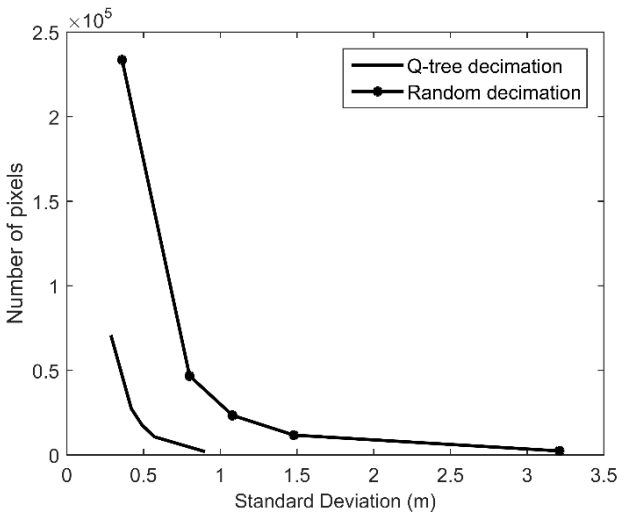
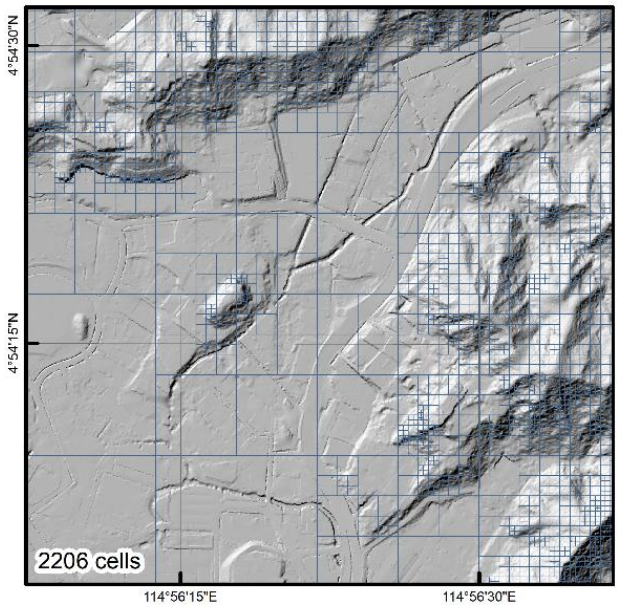
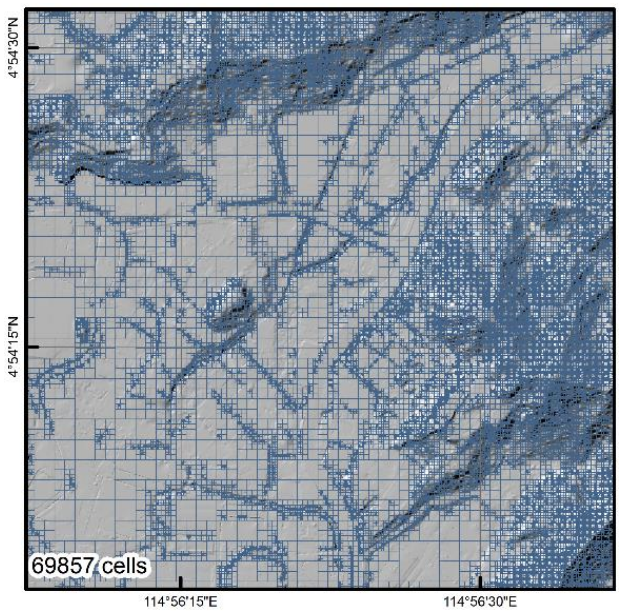


Figure 3. Experimental relationship between the number of pixels used to represent DTMs vs. the standard deviation of the mean elevation difference between the decimated and reference DTMs

Figures 4a and 4b illustrate hillshadow representations of the full resolution DTMs overlaid with the Q-tree with the lowest and highest number of cells, respectively.



a)



b)

Figure 4. Hillshadow DTM Overlaid with the Q-Tree Containing 2206 cells a) and 69,857 cells b)

4. CONCLUSIONS

The aim of this simulation experiment was to propose a potential solution to the BDP in the context of geodata, particularly in the context of datasets used to store information related to topography. The investigated approach was Q-tree surface partitioning with a fixed-σ criterion. The results clearly indicated that the Q-tree significantly outperforms the random data decimation approach in terms of the number of pixels needed to achieve a given accuracy level of the resulting DTM as well as in terms of the fidelity of the results. The latter is clearly illustrated by Figure 4b: the narrow topography features, presumably channels, streams, etc., that are associated with the larger slopes are covered by much smaller cells. The current implementation of Q-tree partitioning with the fixed-σ criterion does not automatically merge neighbouring leaves belonging to different branches of the Q-tree even if the fixed-σ condition is satisfied. This result warrants further studies regarding Q-tree representations of geodata not only related to topography but also to other types of geodata.

ACKNOWLEDGEMENTS

The authors are grateful to the Management of Soartech Systems Sdn Bhd, Brunei Darussalam for providing LiDAR data over the AOI free of charge.

REFERENCES

Becek, K., 2008. Investigating error structure of Shuttle Radar Topography Mission elevation data product. *Geophysical Research Letters*, 35(15).

Becek, K., 2012. A Fixed-σ Digital Representation of a Random Scalar Field. In: Kenyon S., Pacino M., Marti U. (eds) Geodesy for Planet Earth. IAG Symposia, 136. Springer, Berlin, Heidelberg.

Becek, K., 2014. Assessing Global Digital Elevation Models Using the Runway Method: The Advanced Spaceborne Thermal Emission and Reflection Radiometer Versus the Shuttle Radar Topography Mission Case. *IEEE Trans. GRS*, 52(8).

de Berg, M., Cheong, O., van Kreveld, M., Overmars, M. H., 2008. *Quadtrees Non-Uniform Mesh Generation*. Computational Geometry Algorithms and Applications (3rd ed.). Springer-Verlag.

Finkel, R. A. Bentley, J. L., 1974. Quad Trees a Data Structure for Retrieval on Composite Keys. *Acta Informatica*. 4(1), pp. 1–9.

Jianping H., Waibhav D. Tembe, W.D., Dougherty, E.R., 2009. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3), pp. 409-424.

Samet, H., 1984. *The Quadtree and Related Hierarchical Data Structures*. ACM Computing Surveys (CSUR), 16(2), pp.187-260.