DENDROGRAM CLUSTERING FOR 3D DATA ANALYTICS IN SMART CITY

S. Azri*, U. Ujang , A. Abdul Rahman

3D GIS Research Lab, Department of Geoinformation, Faculty of Built Environment and Survey, 81310 UTM Johor Bahru, Malaysia - (suhaibah, mduznir, alias)@utm.my

KEY WORDS: Smart City, Dendrogram Clustering, 3D Spatial Database, 3D GIS, Data Analytics, Data Structure

ABSTRACT:

Smart city is a connection of physical and social infrastructure together with the information technology to leverage the collective intelligence of the city. Cities will build huge data centres. These data are collected from sensors, social media, and legacy data sources. In order to be smart, cities needs data analysis to identify infrastructure that needs to be improved, city planning and predictive analysis for citizen safety and security. However, no matter how much smart city focus on the updated technology, data do not organize themselves in a database. Such tasks require a sophisticated database structure to produce informative data output. Furthermore, increasing number of smart cities and generated data from smart cities contributes to current phenomenon called big data. These large and complex data collections would be difficult to process using regular database management tools or traditional data processing applications. There are multiple challenges for big data, including visualization, mining, analysis, capture, storage, search, and sharing. Efficient data analysis mechanisms are necessary to search and extract valuable patterns and knowledge through the big data of smart cities. In this paper, we present a technique of three-dimensional data analytics using dendrogram clustering approach. Data will be organized using this technique and several output and analyses are carried out to proof the efficiency of the structure for three – dimensional data analytics in smart city.

1. INTRODUCTION

In this few years, 'Smart City' and 'Big Data' are the most hype and buzzwords in business industry, government organizations and even in academia field. According to (Mohanty et al., 2016), smart city can be defined as a place where network, service and infrastructure are more efficient and sustainable with the aid of digital information and advanced technology. The explosive growth of Information and Communication Technology (ICT) has become a pace setter to transform a city to a smart city. Smart city provide a benefits to its inhabitants with a greener, safer, faster and friendlier environment.

During the process and plan of smart city, data are gathered and collected from sensor, social media, and legacy data sources for data analysis to identify infrastructure that needs to be improved or to predict disease outbreak. Data that are produced from smart city plan contributes to current phenomenon called big data. Big data is a large and complex data collection that is difficult to process using regular database management approach. It requires special data handling and sophisticated data structure. Recent reviews from (Hashem et al., 2016) three smart cities; Helsinki, Stockholm and Copenhagen shown that the amount of data generated using Internet of Things (IoT) technologies are huge. For example, 1030 databases were set up in 2013 by The Helsinki Region Infoshare Project to cover a wide range of urban phenomena, such as transport, economics, conditions, employment, and well-being.

Big data are worthless in a vacuum. Its potential value is unlocked only when leveraged to drive decision making. To enable such evidence-based decision making, organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights. According to (Labrinidis and Jagadish, 2012), there are several process to extracting insights from big data. This process can be viewed in Figure 1. There are five stages that are categorized into two main processes; data management and analytics. In the data management phase, technologies are used to acquiring and storing the data. This is important for data preparation and retrieval for further analysis. On the other hand, analytics refers to techniques used to analyse and acquire intelligence from big data. There are a few techniques for big data analytical. The techniques could be used for both structured and unstructured data such as text analytics, audio analytics, video analytics, social media analytics and predictive analytics (Gandomi and Haider, 2015).

* Corresponding author

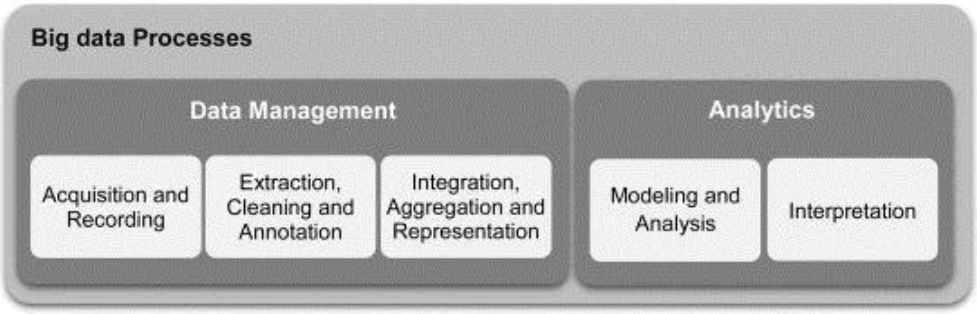


Figure 1. Big data processes (Labrinidis and Jagadish, 2012)

The application of big data in a smart city has many benefits and challenges, including the availability of large computational and storage facilities to process streams of data produced within a smart city environment. In this study, we aim to offer another alternative technique for 3D big data analytics in smart city. Thus, this paper is motivated by the current availability of smart devices that generate large heterogeneous datasets every day and the processing challenges that must be addressed to increase citizens' quality of life and make their cities sustainable. The rest of this paper is organized as follows: problems and motivation regarding the big data of smart city and smart city challenges are discussed in the next section. In Section 3, the concept of the proposed method is explained with its implementation. Section 4 presents the analysis and results of the experiment. Finally, the conclusions are presented in Section 5.

2. RESEARCH PROBLEMS AND MOTIVATIONS

In this section background to the problem on big data is discuss and review. From the problem, proposed technique for 3D data analytics is discuss in the next section.

2.1 Existing Data Structure for Big Data Handling in the Database

In this research, the focus is concentrated on big data processes with a case study of smart city. With the evolution of computing technology, immense volumes can be managed without requiring supercomputers and high cost. Many tools and techniques are available for data management such as Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort (Chen et al., 2014). However, special tools and technologies are required so that it can store, access, and analyse large amounts of data in near-real time. This is because big data are different from the traditional data and cannot be stored in a single machine. Commonly used tools and techniques for big data handling are Hadoop, MapReduce, and Big Table. These tools have redefined data management because they effectively process large amounts of data efficiently, cost effectively, and in a timely manner. However, there are some limitations for these tools such as limitation of queries on Hadoop. In a normal relational database, data is found and analysed using queries. Hadoop is not really a database: it stores data and retrieve data out of it, but there are no queries such as SQL involved. Hadoop is more of a data warehousing system.

While analysing big data using Hadoop has lived up to much of the hype, there are certain situations where running workloads on a database may be the better solution. For example, dealing with structured data that is owing to the fact it is in large volumes that can be entered, stored, queried, and analysed in a simple and straightforward manner, this type of data is best served by a database. In cases where organizations rely on time-sensitive data analysis, a traditional database is the better fit. That's because it can offer shorter time-to-insight the datasets. With regards to the main aim and case study of this research, most of the data generated and used for smart city application is from structured data lead to the option of using database. However, storing and information extraction in database require a new efficient approach. Therefore, this research introduces a new approach of big data handling of smart city in 3D. The approach is efficient to store and retrieve data to acquire the informative output.

To process a huge volume of 3D data in database, a specific 3D data structure is needed for data storing, retrieval and analytics. Current approach in database is by using data constellation to arrange data in memory space. There are two well-known 3D data structure used in database: Octree and 3D R-Tree (Guttman, 1984). Octree is an extended version of two-dimensional quad-tree to three-dimensional. Octree data structure internal node has exactly eight children. Octrees are most often used to partition a three-dimensional space by recursively subdividing it into eight octants. It is widely used for 3D graphics and 3D games application. According to (Keling et al., 2017), the algorithm of Octree is simplified by exploiting recursive nature of Octree, but the drawback is it require long tree traversals. Besides that, voxelization require more storage on memory space.

3D R-Tree seems to be the most promising data structure to be used in database. In fact, it has been widely used in commercial database software such as Oracle. However, the transition of R-Tree to 3D had increase the overlap among node and requires more storage. This would lead to repetitive data and multipath query among node (Azri et al., 2015). Although an effort has been made to improve the structure of 3D R-Tree, application using large amount of data still faces low data retrieval efficiency. Recent review from (Azri et al., 2013) proposed to merge multiple indexing methods to form hybrid data structure. Looking forward to a smart city, there is a need to design a specific data handling or structure to enable computer and system to analyse spatial big data for intelligent decision making. Thus, in this paper, we propose a 3D data structure to constellate big data of 3D smart city data into spatial database.

3. 3D DATA CONSTELLATION IN SMART CITY

In this section the proposed 3D data structure is introduced. The construction and development of the proposed structure is the most important part. All of the information will be accessed through this structure prior to data analytical.

3.1 3D Data Constellation for Efficient Data Retrieval

Classified and Clustered Data Constellation (CCDC) is a 3D data structure that constellate 3D spatial data into spatial database (Azri et al., 2016). The data structure is designed and developed based on two main filters; classification and clustering and works based on hierarchical tree concept. The classification phase will classifies each of spatial objects into a group based on its theme or type. For instance classification based on zoning theme such as, retail, housing or industrial. Then, each object in each group of classification will be clustered using clustering algorithm. In (Azri et al., 2016) the clustering processes are based on k-means++ crisp clustering algorithm by (Arthur and Vassilvitskii, 2007). *k*-means++ introduced the approach of careful seeding to improve the k-means algorithm. By using this approach, initial seeds are defined and the remaining objects are then clustered based on the nearest distance to the initial seeds. This algorithm has proven to yield improvements in terms of accuracy with respect to original algorithm.

The results from classification and clustering phases are then mapped into hierarchical tree structure. Data will be retrieved by traversing the tree structure from its parent node to its child. CCDC data structure offer a very minimal percentage and coverage area among nodes which is one of the requirement for efficient data retrieval from the database. However, still CCDC could not achieve the zero overlap among nodes and we believe that the construction of CCDC structure is time consuming due two group filter which are classification and clustering.

3.2 Dendrogram Clustering

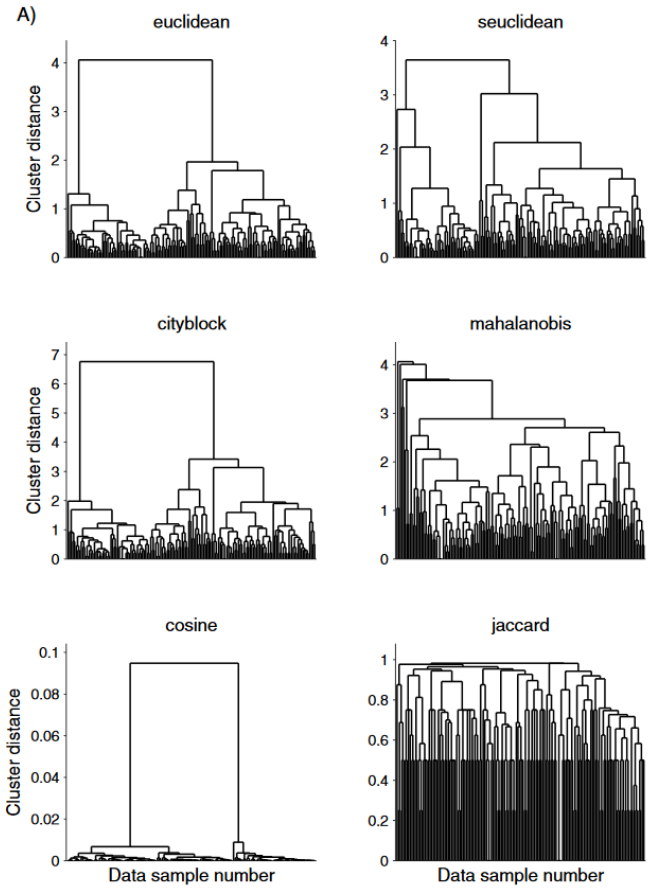
Dendrogram clustering is also known as hierarchical clustering algorithms or HCA. This clustering actually falls into two categories top-down or bottom-up. Bottom-up algorithm categorize each data or object as a single cluster and then merge with its pair until all clusters are merged and become a single cluster. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the objects, the leaves being the clusters with only one sample. Another category is using top-down algorithm or known as divisive. The approach is a bit different from top-down where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In this study, dendrogram clustering is constructed based on bottom up algorithm. The algorithm of dendrogram clustering can be seen as follows and the structure of dendrogram clustering using different distance metric can be seen in the Figure 2.

Algorithm 3.1	**Dendrogram Clustering**
Input:	X Data Points
Output:	X Clusters
1.	**create** single cluster for each X points
2.	**select** distance metric that measures the distance between two cluster
3.	**combine** two clusters into one with a condition smallest distance metric (nearest neighbour)
4.	**repeat** step 3 until reach root of the tree (one cluster which contain all data)
5.	**list** all clusters

One of the advantages of using dendrogram clustering is, it does not require number of clusters. Besides that, the algorithm is not sensitive to the choice of distance metric where it can work equally well with other clustering algorithms. These advantages of hierarchical clustering come at the cost of lower efficiency, as it has a time complexity of $O(n^3)$, unlike the linear complexity of k-means. Dendrogram clustering algorithm is based on a distance matrix that has to be kept in memory. Thus, the distance matrix is symmetric which need memory scales as $\frac{N(N-1)}{2}$. Besides that, average link clustering scales as N3 in time, because for each cluster agglomeration, the algorithm searches through $\frac{N(N-1)}{2}$ cluster dissimilarities in order to determine the pair of most

similar clusters to merge, and the algorithm works through N−1 iterations. However, to overcome this issue the structure can be speed up with cluster seeding as presented in (Embrechts et al., 2013). From their study, *k*-means is applied as a cluster seeding to speed up the performance.

Figure 2. Different dendrograms are produced with



different distance metrics (Fisher, 1936).

3.3 Clustering the Smart City using Dendrogram

Buildings in a smart city are blended with two or more residential, commercial, cultural, institutional, and/or industrial uses. This type of development is also known as mixed-use development. Mixed use is one of the ten principles of Smart Growth, a planning strategy that seeks to foster community design and development that serves the economy, community, public health, and the environment. Mixed-use zoning allows for the horizontal and vertical combination of land uses in a given area. Commercial, residential, and even in some instances, light industrial are fit together to help create built environments where residents can live, work, and play. Figure 3 shows the vertical and horizontal mixed-use development in urban area.

Figure 3. Vertical and Horizontal mixed-use development

To manage this type of development theme in a smart city requires a specific approach than usual. In CCDC data structure, classification will be done at the early stage prior to the clustering process. Building units with the same business such as retail, office or accommodation will be grouped and bounded in a parallelepiped. However, in this study units with different theme will be appeared as different categories and will only be merged before the final grouping. In other words, objects will be clustered with their same categories and these categories will be combined to produce root of the tree. The tree structure is created along the way of clustering process. Unlike CCDC data structure, classification and clustering processes need to be done prior to the creation of the tree.

All information and parameters related to the building in a smart city are stored in the database using dendrogram clustering. The search operation is deployed from the parent node of tree structure and dives in into the group cluster and retrieves the leaf node. Each record in the structure is identified using key identifier. The key identifier consists of unique IDs of group cluster and the building. Algorithm 3.2 presents the dendrogram clustering for processing different types of building in smart city and Figure 4 explain the tree structure for a few 3D buildings sample.

Algorithm 3.2	**Dendrogram Clustering for 3D Building in Smart City**
Input:	3D Building with different type of urban mixed-use theme i.e. Retail, Office or Residential
Output:	Dendrogram
1.	**create** single cluster for each *n* points in x theme
2.	**select** distance metric that measures the distance between two cluster
3.	**combine** two clusters into one with a condition smallest distance metric (nearest neighbour)
4.	**repeat** step 1 to 3 for all data until it become one cluster for one theme
5.	**Combine** all theme to become root tree
6.	**display** tree

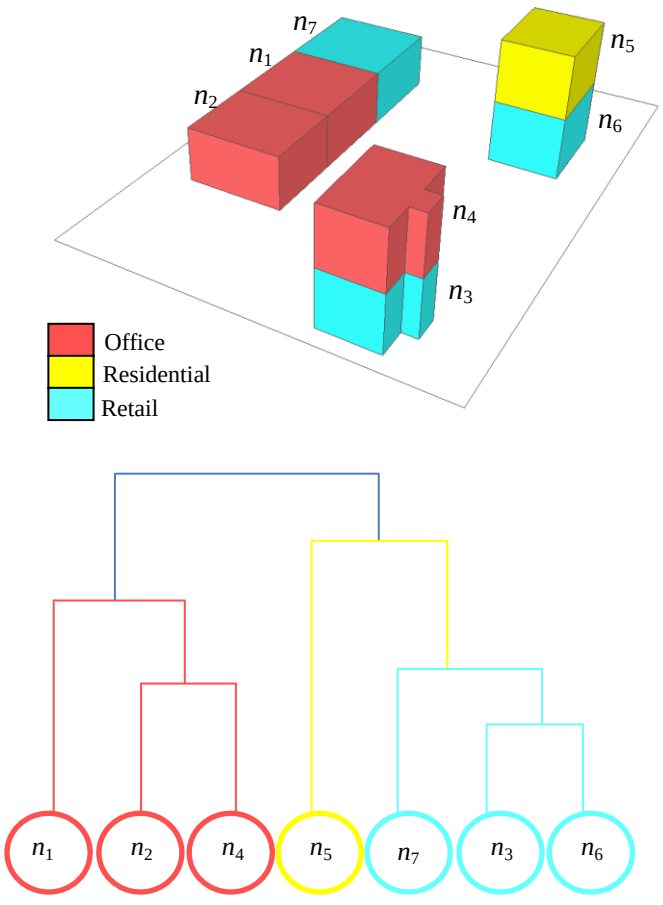


Figure 4. Dendrogram tree for 3D buildings.

From Figure 4, several building with a few units in the building is shown with a different business theme; Office, Residential and Retail. By using dendrogram clustering, these units are organized in a tree structure and grouped with the same business theme.

4. EXPERIMENT AND ANALYSIS

Several experiments and analyses are performed in order to verify the proposed 3D data structure. The experiment is done to retrieve records from database and compare the performance of dendrogram clustering with existing structure.

4.1 Retrieving Records

Records are retrieved from the database using Structured Query Language (SQL). From the statement, the structure specifies the records using its unique identification from cluster ID and then the record can be retrieve directly from the cluster.

Result from Figure 5 show 11 records of roof that are belong to a single building in a smart city. This record reported total incident radiation and absorption of urban heat island phenomenon in that city.

4.2 Page Analysis

An experiment is conducted to evaluate the average number of page accessed or Input/Output (I/O) incurred by the proposed 3D data structure. The test is evaluating the effect of retrieved records from large tuples. To test the performance of query operation using proposed 3D data structure, 1,000 000 records of buildings are used to evaluate the average number of page access. Figure 6 illustrates the I/O cost for finding *k* number of records with and without the 3D data structure. As the figure shows, the proposed data structure offer minimal page accessed. Without the structure the number of accessed page is high due to multipath query and repetitive data entry.

4.3 Response Time Analysis

The main objective in this study is to improve the data retrieval efficiency from the database. Thus, a test is performed to analyse the query response time of the proposed structure. From the tests, *k* number of objects will be retrieved and the time retrieval is measured and recorded in millisecond (ms). 1,000 000 buildings location are populated in the database. Query is performed to retrieve different number *k* of objects from the database. Then, the results are plotted in a graph as presented in the Figure 7. From the result, the proposed 3D data structure outperforms the response time compared to non-constellated data. The response time is 50% to 60% faster compared to non-constellated data.

Object ID	Object Type	Orient.(°)	Tilt(°)	Total Incident Radiation Wh/m2	Total Absorbed Radiation Wh/m2
0	Roof	111.39	-50.72	4490.943	347379.344
1	Roof	111.39	-50.71	4502.512	348274.094
2	Roof	111.39	-52.79	4545.021	351562.406
3	Roof	111.39	-52.8	4553.471	352215.969
4	Roof	-68.61	-20.23	16498.121	1219852.75
5	Roof	-68.61	-20.24	16437.533	1220982.875
6	Roof	-53.61	-20.25	16585.582	1215582.125
7	Roof	-53.6	-20.21	16617.109	1217597.875
8	Roof	-38.6	-20.21	15681.85	1133372
9	Roof	-38.61	-20.18	15733.109	1128865.875
10	Roof	141.39	-52.8	4541.833	351315.625
11	Roof	141.38	-52.79	4544.763	351542.25

Figure 5. Retrieved records from the database.

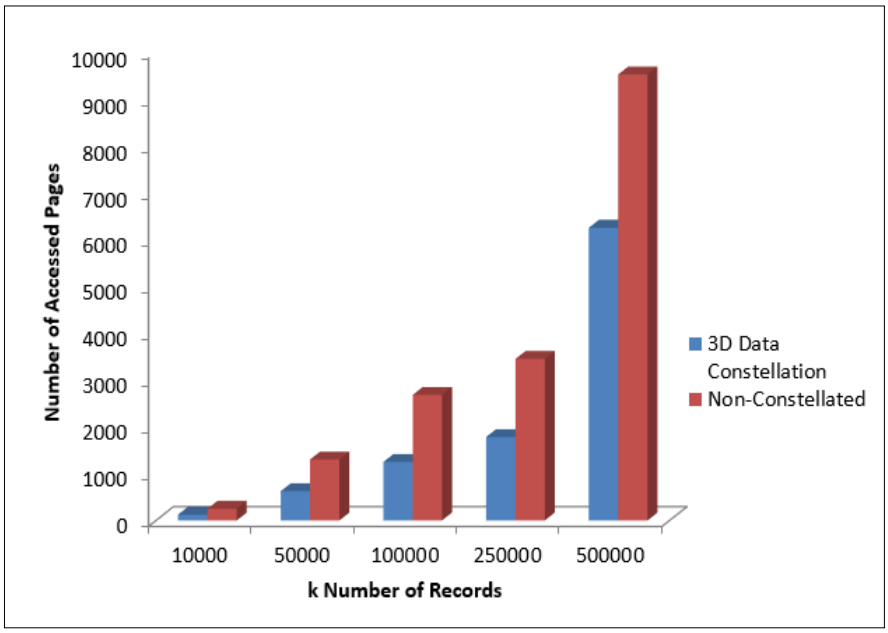


Figure 6. I/O vs *k* number of objects.

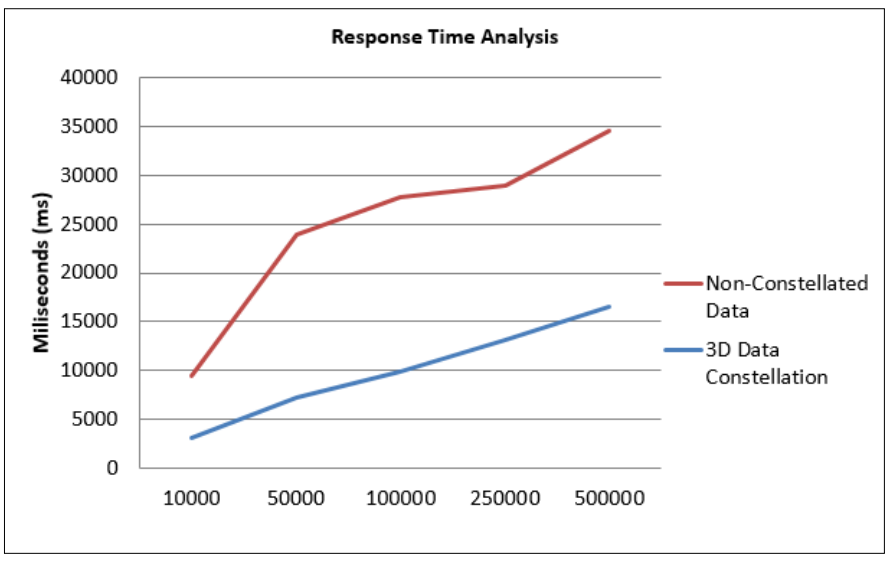


Figure 7. Response time analysis for *k* number of objects retrieval.

This paper proposed 3D dendrogram clustering to produce hierarchical tree structure for data for data retrieval and analytics. The structure is constructed based on dendrogram clustering. The clustering grouped the object with the same features and then groups the object under the root tree. The datasets are retrieved based on specific IDs for each object. The implementation of *k*-means algorithm is used as a cluster seeding to speed up the tree creation. This is due to the lower efficiency, as it has a time complexity of O (n^3). Based on the comprehensive tests and analyses of the proposed structure, results and findings are discussed as follows.

The first test is to prove its ability of the structure in retrieving records from the database. From the test, it is successfully shown that the structure is able to retrieve information of radiation and solar absorption on the roof of one building. The last two experiments were performed in order to measure the efficiency of the structure. Based on page analysis and response time analysis, it is proven that the structure could perform better data analysis with low access to page. We strongly believe that the data structure will be benefitted to planner and spatial professional and aid them to perform better data analytics for smarter cities.

ARTHUR, D. & VASSILVITSKII, S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007. Society for Industrial and Applied Mathematics, 1027-1035.

AZRI, S., ANTON, F., UJANG, U., MIOC, D. & RAHMAN, A. A. 2015. Crisp Clustering Algorithm for 3D Geospatial Vector Data Quantization. *In:* BREUNIG, M., AL-DOORI, M., BUTWILOWSKI, E., KUPER, P. V., BENNER, J. & HAEFELE, K. H. (eds.) *3D Geoinformation Science: The Selected Papers of the 3D GeoInfo 2014.* Cham: Springer International Publishing.

AZRI, S., UJANG, U., ANTON, F., MIOC, D. & RAHMAN, A. A. Review of Spatial Indexing Techniques for Large Urban Data Management. 2013 2013.

AZRI, S., UJANG, U., CASTRO, F. A., RAHMAN, A. A. & MIOC, D. 2016. Classified and clustered data constellation: An efficient approach of 3D urban data management. *ISPRS Journal of Photogrammetry and Remote Sensing,* 113**,** 30-42.

CHEN, M., MAO, S. & LIU, Y. 2014. Big Data: A Survey. *Mobile Networks and Applications,* 19**,** 171-209.

EMBRECHTS, M. J., GATTI, C. J., LINTON, J. & ROYSAM, B. 2013. Hierarchical clustering for large data sets. *Advances in Intelligent Signal Processing and Data Mining.* Springer.

FISHER, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics,* 7**,** 179-188.

GANDOMI, A. & HAIDER, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* 35**,** 137-144.

GUTTMAN, A. 1984. R-trees: a dynamic index structure for spatial searching. *SIGMOD Rec.,* 14**,** 47-57.

HASHEM, I. A. T., CHANG, V., ANUAR, N. B., ADEWOLE, K., YAQOOB, I., GANI, A., AHMED, E. & CHIROMA, H. 2016. The role of big data in smart city. *International Journal of Information Management,* 36**,** 748-758.

KELING, N., MOHAMAD YUSOFF, I., LATEH, H. & UJANG, U. 2017. Highly Efficient Computer Oriented Octree Data Structure and Neighbours Search in 3D GIS. *In:* ABDUL-RAHMAN, A. (ed.) *Advances in 3D Geoinformation.* Cham: Springer International Publishing.

LABRINIDIS, A. & JAGADISH, H. V. 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment,* 5**,** 2032-2033.

MOHANTY, S. P., CHOPPALI, U. & KOUGIANOS, E. 2016. Everything you wanted to know about smart cities: The Internet of things is the backbone. *IEEE Consumer Electronics Magazine,* 5**,** 60-70.