COMPARATIVE ANALYSIS OF DATA STRUCTURES FOR STORING MASSIVE TINs IN A DBMS

K. Kumar, H. Ledoux, J. Stoter

3D Geoinformation, Delft University of Technology, Delft, The Netherlands - (k.kavisha, h.ledoux, j.e.stoter@tudelft.nl)

Commission II, WG II/2

ABSTRACT

Point cloud data are an important source for 3D geoinformation. Modern day 3D data acquisition and processing techniques such as airborne laser scanning and multi-beam echosounding generate billions of 3D points for simply an area of few square kilometers. With the size of the point clouds exceeding the billion mark for even a small area, there is a need for their efficient storage and management. Storing billions of 3D points is currently possible which is confirmed by the initial implementations in Oracle Spatial SDO_PC and the PostgreSQL Point Cloud extension. But to be able to analyse and extract useful information from point clouds, we need more than just points i.e. we require the surface defined by these points in space. There are different ways to represent surfaces in GIS including grids, TINs, boundary representations, etc. In this study, we investigate the database solutions for the storage and management of massive TINs. The classical (face and edge based) and compact (star based) data structures are discussed at length with reference to their structure, advantages and limitations in handling massive triangulations and are compared with the current solution of PostGIS Simple Feature. The main test dataset is the TIN generated from third national elevation model of the Netherlands (AHN3) with a point density of over 10 points/m². PostgreSQL/PostGIS DBMS is used for storing the generated TIN. The data structures are tested with the generated TIN models to account for their geometry, topology, storage, indexing, and loading time in a database. Our study is useful in identifying what are the limitations of the existing data structures for storing massive TINs and what is required to optimise these structures for managing massive triangulations in a database.

1. INTRODUCTION

Airborne LiDAR technology is nowadays extensively used to collect 3D information in the form of dense point clouds, quickly and with great accuracy. AHN (Actueel Hoogtebestand Nederland), the national elevation dataset of the Netherlands, contains billions of points with a point density of 6-10 points/m². With the size of such point clouds exceeding the billion mark even for the smaller areas, difficulties arise in storing and managing these datasets. The existing DBMS solutions such as the Oracle Spatial SDO_PC package (Ravada et al., 2009) and the Postgresql Point Cloud extension (Ramsey, 2013) provide only for the storage of point clouds i.e. unconnected points that are samples of the surface representing the Earth. However, in order to analyse and extract useful information like slope/aspect, viewed analysis, watersheds modellings, delineation of basins, etc. from a point cloud, we require more than just points: we must be able to reconstruct and manipulate the surface defined by these points. Surface defines the spatial relationships between the disjoint points in 3D space. The most common method for representing a surface in GIS is the construction of a 2.5D structure such as TIN (Triangulated Irregular Network) (Peucker et al., 1978). Practitioners often think of grids as the simplest way to represent a surface. In fact, it may be the best model to depict the real world scenarios, but this possibility depends on the potential of the grids rather than their current use (Fisher, 1997). In this study, we investigate the database solutions for the storage and management of massive TINs. Here, the term “massive” refers to extremely large size which is difficult to manipulate and manage with main memory. The more complex representations of TINs take into account the attributes and constraints associated with the connecting elements. The local density of the points in a TIN can be adjusted in accordance with the variations in the height of original terrain e.g. TINs naturally represent areas of detailed relief with denser triangulation than areas with a smoother relief (Kumler, 1994). While the enormous size of the point clouds is already an issue, storing the derived TIN significantly increases the storage requirements since the number of triangles generated is about twice the number of points taken into account for triangulation (De Berg et al., 2000). AHN2 point cloud dataset has over 640 billion points. The number of triangles generated during triangulation is thus around 1.3 trillion.

Storing massive TINs in the database is much more complicated than storing point clouds since storage of TIN not only requires storing its geometry but the topology as well. The requirement is to reduce the redundancy of explicit representations while still supporting mesh traversal and modifications to dynamically maintain the mesh in the event of local updates like add/delete a vertex/triangle, edge flipping, etc. Essentially, the ideal characteristics of storing TINs in a database include low storage costs for geometry, topology and other attributes, simplicity, easy modifications/updates, faster access to the adjacency information and scalability. Solutions that handle the issue of massive point clouds such as tiling are less viable in the case of TINs. When the dataset is too large to be processed within the available main memory, it is split into smaller parts (called tiles). The processing is done incrementally on the generated tiles. For the analysis of TINs, the main memory of the computer plays a key role in deciding the maximum size of the dataset that can be processed (Isenburg et al., 2006). If the dataset exceeds the memory size, then the swapping of data between the disk and the main memory starts, thereby increasing the processing time. This process of tiling cannot be extended to TINs without breaking its topology (Fig. 1) unless there is addition of explicit points on the tile borders (called Steiner points) causing changes in TIN topology (Fig. 2).
2. RELATED WORK

2.1 TIN Structure

A TIN is a network of non-overlapping triangles formed by the interconnection of irregularly spaced points (Kumler, 1994). They are mostly constructed with Delaunay triangulation but this is not always the case. For a given set of points, different triangulations can be constructed (Rippa, 1990, Dyn et al., 1990). A Delaunay triangulation (DT) of a point set S (Fig. 3) is defined as a triangulation of S such that no point in S lies inside the circumsphere of any other triangle in the triangulation (De Berg et al., 2000). DT maximizes the minimum angle among all possible triangulations to avoid long and skinny triangles.

But everytime the usual inputs for TIN generation are not merely a set of vertices. If the point set is associated to some constraints (segments, polygons, etc.) then a Constrained Delaunay triangulation (CDT) can be constructed. A CDT is similar to DT but every input segment appears as an edge of the triangulation (Shewchuk, 1996). CDT requires that the vertices of every triangle are visible to each other i.e. no input segment (or a part of it) lies between its any two vertices and the circumcircle of each triangle contains no vertices that are visible from the interior of the triangle (Shewchuk, 1996). Fig. 4 and 5 show an example of CDT.

There exists several data structures for representing TINs in memory but only a little has been done so far for their database implementations. We provide here a comparative analysis of the existing TIN data structures. The classical (face and edge based) and compact (star based) data structures are discussed at length with reference to their structure, advantages and limitations in handling massive triangulations. Our aim is to compress these structures while maintaining the topological relationships. In this paper, we explain how to implement these structures in a database. These data structures are tested with a few real world datasets and we report on their geometry, topology, storage, indexing, and loading time in a database. The study is useful in deriving what are the limitations of the existing data structures in storing massive TINs and what is required to optimise these structures for managing massive triangulations in a database.
Most of the TIN data structures stores triangulations as an array of vertices or as references to the vertices forming the triangles, thereby capturing the mesh geometry. OGC Simple Feature (OGC, 2011) is a classical example of such data structures and is supported by almost all the spatial databases. It stores each triangle as a linear ring of vertex coordinates with last vertex = first vertex as depicted in Fig. 6. The size of the database increases considerably with the repeated storage of vertex information for each triangle and creation of complex GIST spatial index on the triangle geometry. The average degree of a vertex in a 2D Delaunay triangulation is 6, given that the vertices follow Poisson distribution (Okabe et al., 2009). This suggests that on an average each vertex is stored 6+6/3 = 8 times. In general terms, each vertex is stored \((1+t/3)\) times in the database with Simple Feature, where \(t\) is the number of incident triangles on a vertex. Here the operand \(t/3\) suggests the fact that given vertex can be the first vertex (and thereby the last vertex) in \(t/3\) number of incident triangles. It does not store explicitly the adjacency relationships between the triangles, which is helpful for the traversal of the triangular mesh.

Similarly, what we refer here as the Triangle data structure stores a triangulation as references to the \(1Ds\) of three vertices forming the triangle (Fig. 7). The coordinates are not repeated here as in the Simple Feature. Oracle Spatial SDO_TIN (Ravada et al., 2009) uses Triangle for storing TINs. The vertices are stored as SDO_PC type and the triangles are stored as references to their three vertices. A notable point of this approach is that the indexing is performed at the block level and not at the vertex or triangle level. The limitation of this approach is that it lacks information about the adjacency and incidence relationships between the triangles and the vertices are not stored explicitly. Also, it suffers from the problem of redundancy, when the triangles span more than one block (Fig. 1). Maintaining the topology of the TIN is by far the most expensive operation.

A more convenient approach is to store the triangles along with their adjacency information (here we refer to it as Triangle+).

Likewise, the edge based data structures namely, the Half-edge (Mäntylä, 1987) or DCEL (Doubly Connected Edge List) (Muller and Preparata, 1978) represents each edge as a composition of two half-edges (counter clockwise and opposite) in a slightly non-intuitive fashion (Fig. 9). Apart from storing the adjacency relationships (i.e. references to previous edge \(p(e)\), next edge \(n(e)\) and the incident face \(f(e)\)), DCEL maintains references to the incident half edges \(e\) and \(e'\) and to the bounding vertex \(v\) (De Berg et al., 2000). While the number of edges \(e\) in a triangulation is about thrice the number of vertices \(n\), this results in \(3e+n\) references for the mesh with \(19rpt\) or \(6rpt\). Other variations like the Winged-edge (Baumgart, 1975) stores eight references per edge of the triangle. The storage requirements of these data structures is high which is accompanied by the large size of their spatial index.

Apart from the classical representations, several compact data structures for the triangular meshes have been described to reduce the number of references required and the memory requirements while still ensuring an efficient implementation in terms of run time and mesh traversal operations. Blanford et al. (2005) describes a compressed, pointer less, star based data structure for compactly representing the triangular meshes in two and three
The main advantage of the star data structure is that it does not require creation of complex spatial index like giST. The indexing is computed on-the-fly. Each incident triangle at the bounding vertex is represented as an ordered list (counter-clockwise) of vertex labels \( v_1, v_2, v_3, \ldots, v_j \) forming the link (Fig. 9). The triangles are computed on-the-fly. Each incident triangle at the bounding vertex is represented by \( v \) and the two consecutive vertices \( v_i \) in the list. On an average, each vertex in the star stores reference to 6 neighboring vertices, which results in a storage cost of \( 6r pv \).

The advantage of the star data structure is that it does not require creation of complex spatial index like giST. The indexing is done at vertex level with a simple B-tree.

For practical use, we attempt to create a data structure which can be built on-the-fly, simultaneously with mesh generation-processing. Table 1 highlights the comparison between the existing data structures for the triangular meshes.

### 3. TINS in DBMS

Traditionally, DBMS were used for handling administrative and other voluminous data but now they have evolved to integrate the spatial component. At present, database systems like Oracle, PostGIS, IBM DB2, Ingres, Informix, MySQL, etc. provide support for spatial data types, spatial indexing and other extended functionalities. Storage and analysis of spatial data can be done with SQL queries. Most of these database systems offer 2D data types (generally point, line, and polygon). Only two database implementations for TIN type are seen so far, namely Oracle SDO_TIN and OGC Simple Feature Access model in PostGIS. OGC Simple Feature Access model (OGC, 2011) specifies the geometrical model to be followed for representing geographical objects. It describes a triangle as a polygon with 3 non-collinear and distinct points and not having any interior rings or holes. A TIN is described as a contiguous collection of triangles (a polyhedral surface) which share common edges. In 3D, with the inclusion of z-values (height), they are referred as TriangleZ and TINZ respectively. At present, the support for storing TriangleZ and TINZ in the database is provided by PostGIS v2. The WKT (Well Known Text) for representing TriangleZ and TINZ is:

- **Trianglz** \(((0\ 0\ 0, 0 \ 1 \ 0, 1 \ 1 \ 0, 0 \ 0 \ 0))\) and
- **TINZ** \(((0\ 0\ 0, 0 \ 1 \ 0, 0 \ 0 \ 0), (0 \ 0 \ 0, 0 \ 1 \ 0, 1 \ 1 \ 0, 0 \ 0 \ 0))\).

In Simple Feature, the geometry is stored with a unique ID, bounding box, and WKB (Well Known Binary) representation of OGC Geometry. Fig. 11 describes the WKB format for OGC Simple Feature type.

Here B represents the Byte order (big endian = 0 and little endian = 1), T represents the WKB Geometry type, NR denotes number of rings in the geometry, NP denotes number of points in the geometry and X, Y, Z are the coordinate values of the points.

Fig. 12 describes the structure of Triangle and TIN type in PostGIS. PostGIS TIN type stores only the geometry of TIN and not the topology.

**Oracle SDO_TIN** makes use of Triangle data structure for storing triangulations (Fig. 14). It is an extension of Oracle SDO_PC type with almost same structure and storage scheme but has an additional column to store the TIN objects. The SDO_PC partitions the point cloud into multiple blocks which are uniquely identified by their block IDs and then spatially indexed (Finnegan and Smith, 2010). The SDO_TIN object stores TIN metadata and a reference to the TIN Block table (\(SDO\_TIN\_BLK\)). Fig. 13 explains the structure of SDO_TIN. The TIN is divided into blocks/buckets and is stored in the TIN Block table with points and triangles as BLOBs (Binary Large OBjects). A point BLOB consists of an array of points of size \(d \times 8\)-bytes (\(d = \) dimensions, 2 in our case) along with Bucket/Block ID and Point ID (\(BLK\_JD, PT\_JD\)) of each point. Each triangle BLOB comprises of an array of triangles, wherein each triangle has 3 vertices and each vertex is referred by the pair of Bucket/Block ID and Point ID (\(BLK\_JD, PT\_JD\)). Each row of the SDO_TIN.BLK table can be considered itself as a complete and connected bucket. Moreover, this explicit storage of block ID with each point solves the issue of dividing TIN with bucketing. (Ravada et al., 2009).

The **Oracle SDO_TIN** does not store the topology of TIN. This can be supplemented with the triangle adjacency information as

---

Figures and tables are described in the text for clarity. The provided text includes references to previous work and provides a detailed explanation of the star data structure and its advantages over other indexing methods. The comparison between different database systems for handling TIN data is also discussed, with a focus on the Oracle SDO_TIN and OGC Simple Feature Access models in PostGIS.
<table>
<thead>
<tr>
<th>Data Structure</th>
<th>Type</th>
<th>rpt</th>
<th>Topology</th>
<th>Features</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>PostGIS SF</td>
<td>Face based</td>
<td>4</td>
<td>X</td>
<td>Face as linear ring of vertices</td>
<td>Vertex information redundancy, No topology, Large spatial index, No topology</td>
</tr>
<tr>
<td>Oracle Triangle</td>
<td>Face based</td>
<td>3</td>
<td>X</td>
<td>Triangle as reference to 3 vertices only, Bucketing of vertices, Block level indexing</td>
<td>No topology</td>
</tr>
<tr>
<td>Triangle+</td>
<td>Face based</td>
<td>6</td>
<td>✓</td>
<td>Triangle as reference to 3 vertices only, Explicit topology storage</td>
<td>Triangle level indexing</td>
</tr>
<tr>
<td>DCEL</td>
<td>Edge based</td>
<td>6</td>
<td>✓</td>
<td>Triangles as reference to the half edges</td>
<td>Edge information redundancy, Large spatial index</td>
</tr>
<tr>
<td>Star</td>
<td>Vertex based</td>
<td>3</td>
<td>✓</td>
<td>Triangles as ordered list of vertex labels and are computed on the fly, Vertex adjacency information, More compact, No spatial indexing but simple B-tree indexing at vertex level</td>
<td>No dynamic updates</td>
</tr>
</tbody>
</table>

Table 1: Existing TIN data structures

Figure 12: Triangle and TIN type in PostGIS

seen in the Triangle+ data structure in section 2. Storing triangulation requires a table where each row contains the IDs of the three vertices forming the triangle. Adding the adjacency information requires creation of another table with triangle ID and IDs of three neighbouring triangles sharing common edges with it. Fig. 15 shows the structure of Triangle+ approach in the database. Using this approach, only 6rpt are required to store a single triangle with topology in the database.

Among other implementations, the pgTIN extension for PostgreSQL (Ledoux, 2015) stores TIN as uncompressed stars. It only makes use of a vertex table with each row containing a vertex ID, vertex coordinates and an array of references to the vertices in the star of the vertex as explained in section 2. Fig. 16 describes the structure of stars in a database. Triangles are computed on-the-fly using the explicitly stored topology. Thus, only 3rpt are required to represent a triangle. However, each vertex in the table is indexed, although using a binary tree, the index so
generated is large.

<table>
<thead>
<tr>
<th>STARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ point_id</td>
</tr>
<tr>
<td>+ point_coordinate_x</td>
</tr>
<tr>
<td>+ point_coordinate_y</td>
</tr>
<tr>
<td>+ point_coordinate_z</td>
</tr>
<tr>
<td>+ star (point_id)</td>
</tr>
</tbody>
</table>

Figure 16: Stars UML

Fig. 17 depicts the row structure of the above discussed data structures in PostgreSQL.

![Row structure of data structures](image)

Figure 17: Row structure in SF, Triangle, Triangle+ and Star

4. IMPLEMENTATION

4.1 Construction of TIN

Two filtered subsets and an AHN3 Tile# 37EN/1 (PDOK, 2015) of the massive AHN3 point cloud dataset are utilised as inputs for the construction of TIN. The TIN is generated based on the concept of spatial streaming given by Isenburg (2006). The mesh construction pipeline makes use of spfinalize and sp delaunay2d modules of the blast extension of Lastools. These modules exploit the spatial coherence of the massive point clouds to compute a streaming triangulation. (Isenburg et al., 2006). The TIN is created and stored without any attributes and constraints. The details of the input datasets are given in Table 2. It lists the number of points in each subset, the number of triangle generated in triangulation, the points forming the convex hull of the subset and the average and maximum number of edges incident on a vertex i.e. the average and maximum degree of a vertex in the triangulation.

4.2 Testing

The PostgreSQL/PostGIS DBMS is used for storing the TIN. For the storage of vertex IDs and triangle IDs, 64-bit large range integers i.e. bigint because 32-bit integers limit the dataset to \( 2^{32}/2 \) (i.e. around 2 billion) IDs. The coordinates \((x, y, z)\) of the vertices are stored in three separate columns as 64-bit double precision. For the storage of star of a vertex, variable length integer array is used because the link of a vertex can contain 2 to an infinite number of vertices. To store the triangles, the triangle ID and the IDs (64-bit integers) of its three vertices are stored in three separate columns. The adjacency relationships for the Triangle+ are stored in a similar manner. For the storage of attributes or constraints associated with the triangles, a new column with proper data type must be created. The Python implementations of the data structures are used to populate the PostgreSQL/PostGIS database using the *.smb outputs of spdelaunay2d. The testing is done on Intel(R) Xeon(R) CPU E5-2650 v3 (2.30GHz) running Ubuntu (Linux), Python 2.7.6 and PostgreSQL v9.3.11.

To compare these data structures, the TIN is stored:

1. in the triangle/triangle+/star based representations with B-tree index.
2. in the PostGIS Simple Feature with GiST index.

Table 3 enlists the time taken for the construction and loading of TIN and spatial index of datasets for different data structures. The storage space used in PostgreSQL is given in table 4.

5. RESULTS AND DISCUSSION

From Table 3, it is observed that it takes significantly large amount of time to populate and index the TIN datasets with PostGIS SF i.e. \(~15\) hours for a TIN generated from one tile of AHN3. The minimal time is taken by the Star data structure i.e \(~4\) hours, followed by the Triangle. This is around 4 times less than that of the SF. This can be owed to the construction of complex PostGIS GiST index which took around 14 times more time than indexing stars with a B-tree.

From Table 4, it is clear that GiST is larger than B-tree as it takes the maximum disk space i.e. \(~202\) GB for storing a TIN generated from a single tile of AHN3. Stars take the least space of about 77 GB for the same. Storing TINs in PostGIS SF is memory expensive because of two reasons: First, the triangles are almost twice in number than the points and with vertex information redundancy there is a significant wastage of storage space of (number of triangles * (4+3) * 8) bytes. Second, the bounding box of every triangle is explicitly stored in PostGIS. The memory requirement statistics for the data structures in the database can be seen as:

1. **SF** - (number of triangles * 4 * number of dimensions * 8) bytes
2. **Triangle** - ((number of triangles * 4 + 4) + (number of vertices * 3 + 8)) bytes
3. **Triangle+** - ((number of triangles * 7 + 4) + (number of vertices * 3 + 8)) bytes
4. **Star** - (number of vertices * 4 + number of elements in each link * 4) bytes
6. CONCLUSION

Storage and analysis of massive TINs can be tackled in different ways. One way is to store them in file based systems either as point clouds (*.laz, *.xyz, etc.) or as a collection of triangles (*.obj, *.smb, etc.). But with increasing size of the point clouds and derived TINs, the storage size and requirements are also increased. Therefore, we have investigated in this paper the data structures which can be implemented as DBMS solutions for the storage of TINs. From the tests conducted, it can be concluded that PostGIS SF can be overlooked for the storing TINs as far as memory usage, loading and query times, indexing and topology is concerned. Although storing triangles as references to their vertices is quick, it suffers from lack of topological information when it comes to TIN traversal and updates. Other representations like DCEL seems to be excessive in terms of adding bookkeeping concerns for object references which require a dependency over robust memory management subsystem to work efficiently with massive triangulations. These classical structures are not useful if the mesh contains a large number of unconnected triangles since the neighbouring information depends on the connectivity of triangles. A compact yet fully connected data structure is the requirement for storage efficient representation of TINs. The requirement for storage efficient representations for triangular meshes has resulted in a number of compression methods and light weight data structures like SQuad (Gurung et al., 2011a), Groapper (Luffel et al., 2014), Laced Ring (LR) (Gurung et al., 2011b), Zipper (Gurung et al., 2013), and Tripod (Snoeyink and Speckmann, 1999). These data structures just stores a set of core links from which other links can be inferred. In the testing results, Triangle+ and Star seems to be promising representations for the storage and management of TINs. The future research holds for rigorous testing with the compact representations and comparing the analysis results with that of the Triangle+ and Star. Data structures are also to be implemented with tiling the dataset into non-massive tiles/blocks to see how efficiently the topological relations across the tiles can be inferred, thereby making it possible to traverse across the tiles. The data structures (with and without tiling) are to be tested for their practical performance by implementing spatial analysis functions and triggering spatial queries on the stored data.

ACKNOWLEDGEMENTS

This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

REFERENCES


ACKNOWLEDGEMENTS

This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

REFERENCES


