Big LIDAR Data in Digital Earth: Ways Out of Dead End*

Ilya Rylskiy [0000-0002-7675-4621]

Lomonosov Moscow State University, 119991, Leninskie Gory, b. 1, Moscow, Russian Federation
rilskiy@mail.ru

Abstract. During past 25 years, laser scanning has evolved from an experimental method into a fully autonomous family of Earth remote sensing methods. Now this group of methods provides the most accurate and detailed spatial data sets, while the cost of data is constantly falling, the number of measuring instruments (laser scanners) is constantly growing. The volumes of data that will be obtained during the surveys in the coming decades will allow the creation of the first sub-global coverage of the planet. However, the flip side of high accuracy and detail is the need to store fantastically large volumes of three-dimensional data without loss of accuracy. At the same time, the ability to work with the specified data in both 2D and 3D mode should be improved. Standard storage methods (file method, geodatabases, archiving, etc) solve the problem only partially. At the same time, there are some other alternative methods that can remove current restrictions and lead to the emergence of more flexible and functional spatial data infrastructures. One of the most flexible and promising ways of laser data storage and processing are quadtree and octree-based approaches. Of course, these approaches are more complicated than typical file data structures, that are commonly used for LIDAR data storage, but they allow users to solve some typical negative features of point datasets (processing speed, non-topological spatial structure, limited precision, etc.).

Keywords: LIDAR, Big Data, Digital Earth, Octree.

1 Introduction

Laser scanning, or LIDAR, is one of the youngest types of surveying. A laser shooting device (laser scanner) is a laser range finder that performs a single or group of simultaneous range measurement with simultaneous measurement of beam deflection angles (vertical and horizontal plane). If each of the measured reflections has

* Publication is supported by RFBR grant № 19-07-00844.

accurate time stamps, the data can be synchronized with high-precision inertial navigation systems (INS) and GNSS tools for positioning in global coordinate systems.

In addition to coordinate information, a laser scanner can register the amplitude of the reflected signal, the albedo of the surface reflecting the signal, as well as register more than one reflected signal, classify the signals according to the shape of the reflected pulse, and much more.

The method is rapidly developing since the late 1990s. The number of laser scanning systems produced during this time around the world is (according to various estimates) several thousand units, with the vast majority of these systems released in recent years.

Each of these systems is characterized by the highest data performance. However, most of these systems (due to the extremely high cost) are constantly in operation. All of these LIDARs continuously produce a huge amount of data. However, the task of combining these enormous datasets into services like Google Earth has not yet been resolved even at a conceptual level. Storing petabyte-sized data arrays while preserving the possibility of quick access requires different approaches compared to those accepted now.

2 Defining problems

2.1 Modern types of LIDAR systems

By mid-2020, several main types of laser scanning systems can be distinguished. Roughly all systems can be divided into two large classes, operating on the principles of measuring the time of flight of the signal (TOF-LIDAR), and measuring the phase of the incoming radiation wave generated earlier by the device itself at frequencies substantially lower than the radiation frequency of the scanner itself (phase LIDAR).

Phase-range lasers generate a modulated light signal. When measuring distance, the phase shift of the emitted and received signals is measured. With significant range, the scanning frequency and accuracy are reduced. High performance is achieved by reducing the range to 120, 50 and even 15 meters - while measuring accuracy increases. A number of algorithms exist and have been tested for operation in conditions of ranges exceeding limits mentioned above. At the same time, the accuracy suffers: if at short distances it can be sub-millimeter, then at a distance of 150 m achieving accuracy of 20-30 mm is very problematic.

Lasers that measure the time of flight of a signal theoretically have no reasonable range limitations at all - it is determined only by the power of the laser pulse and the accuracy of beam focusing. However, they have limitations on accuracy. In 2020, the most advanced instrument samples have a range determination accuracy of about 3 mm (over a range of ranges from 5 to 500 m) and up to 25 mm at ranges of up to 6,000 m and more [1].

Each of these LIDAR types also can be divided into subclasses - scanners for static and for kinematic measurements. Systems for static work are usually called terrestrial laser scanners, systems for kinematic measurements (in motion) are usually also divided into sub-classes: airborne laser scanners (for UAVs or manned carriers) and

mobile laser scanners (for working with cars, trains, special carriers or being transported by a person).

Terrestrial systems may not have accurate time stamps and may not be able to be automatically accurately positioned on the surface of the Earth. Kinematic systems, in principle, cannot exist without this feature, and are always positioned first in global coordinate systems (for example, WGS84), and only then are they subject to adjustment procedures.

2.2 Performance of modern LIDAR systems

The performance of all modern kinematic scanning systems is extremely high. The most advanced airborne scanning systems - for example, the Riegl Q1560ii - have a scanning speed of 4,000,000 points per second [2]. The most advanced mobile systems - for example, Riegl VMX RAIL - up to 3,000,000 points per second [3]. Systems of other companies (Leica, Optech, a number of others) have worse characteristics, but their flagship systems also reach 1,000,000 points per second.

Static scanning systems also allow surveying at speeds of over 1,000,000 measurements per second (Riegl VZ2000i). At the same time, as a rule, all systems are also equipped in parallel with cameras synchronized with laser scanners. This makes it possible (with a high-quality calibration of all devices) to assign to each scan point also the RGB brightness attributes of a given surface area. It allows to make laser point cloud colorized.

Thus, all the described systems are capable of creating spatial data of uncompromising accuracy with a speed not previously encountered. It is easy to calculate that in 5 hours of full-time continuous operation, a system with a performance of 2 million points will produce about 36 billion points. In fact, it is about 50 billion, since one impulse usually gives more than one reflection in the presence of vegetation at the territory. If we briefly estimate the data storage capacity of 60 bytes per point, this gives about 3 terabytes of laser scanning data only. And here we did not take into account that aerial images are also usually being associated with laser scanning.

In terms of 1 km2, the amount of data also looks quite impressive. Now it is not uncommon to ensure accuracy of 10 points per 1 m2 and detail of images at 5 cm / pixel. This requires 600 bytes for laser scanning data and approximately 200 bytes for storing the finished orthomosaic in JPG format, for a total of 800 bytes per 1 m2. Or 0.8 gigabytes to store data per 1 km2. To store data on an area of a medium-sized region of the Russian Federation with an area of 100,000 km2, 80 terabytes will already be required. Up to 14 petabytes of space will be required to store data of the above detail on the territory of the Russian Federation (Fig.1).

Taking into account the above factors, it can be noted that the task of storing and processing laser scanning data in large volumes in the near future will become more and more difficult. Solving this problem will require more and more specific approaches, more typical for working with other types of Big Data [4].

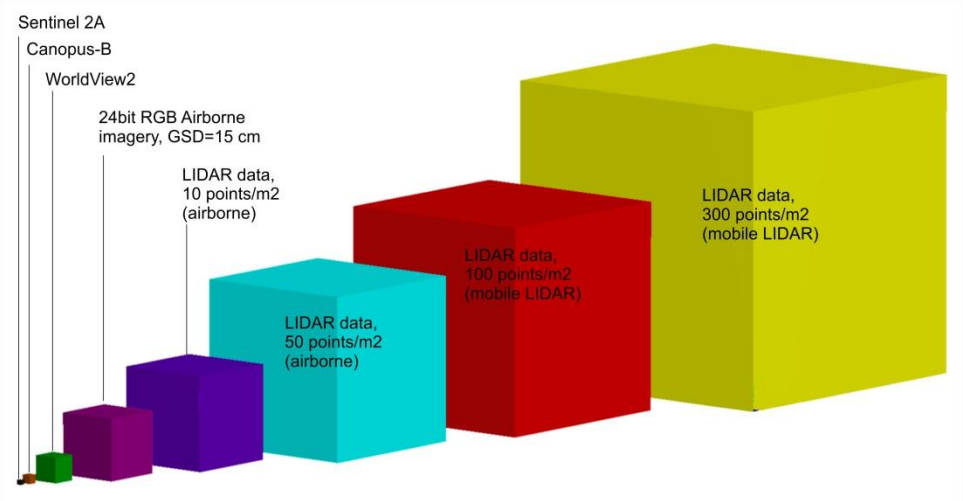

Fig. 1. 3D chart, illustrating the relative volume of spatial data per 1 km2, produced by satellite sensors, airborne photography and LIDAR.

3 Current and perspective methods of LIDAR data storage and processing

3.1 File data storage - inconvenient reality

Despite the growing difficulties in use, the most common method of storing information is to use a file storage system with spatial segmentation. During the implementation of each project, the user creates a system of spatial polygons, each of which corresponds to a certain file with the spatial information stored in it (points of laser reflections, orthomosaics, etc.). These can be regular rectangles (tablets), trapeziums of nomenclature sheets in state patterns, pieces of buffer zones for a linearly extended object, overlapping polygons of complex shape, and other geometric objects. The connection between vector objects (as a rule, this is GIS data with a layered storage system) and physical files on a disk is usually established using a hyperlink or as a text string with a way to data file location.

The disadvantages of this method are quite obvious. First of all, it is needed to store materials in one or several projections without the ability to switch quickly from one projection to another. Projects with such an approach usually cannot be simply combined among themselves, either because of different coordinate systems or different types of “slicing” and just different technical parameters of the survey. It is also not easy to select the optimal size of the “storage element” - a too small storage area requires constant work with a large number of pieces, too large does not allow loading all the data into RAM or reduces the workstation’s productivity.

However, the most unpleasant aspect of such approach is the form of storing laser scanning data in the form of dots itself [5]. Almost all modern data storage formats -

PLY, LAS, FBI, BIN and the like - use a data structure in the form of sequential records for each laser point, each of which contains a time stamp, coordinates, color, amplitude and a number of other parameters. The laser scan data file is thus a gigantic table.

Despite the fact that in this type of record the data is spatial, they are not topological. Even an elementary query like "select points within a radius of X from point Y" will require complete processing of the entire data array in the current data segment. If the point is closer than at a distance X from the border of the current segment, then processing can be stopped and a new segment will need to be loaded (often manually). To avoid such situations, segments are often made overlapping [6], which does not facilitate the operation of algorithms and requires even more disk space.

This is a critical issue with a file-based segmented approach to data storage. As the size of the segment decreases, its severity increases. With very dense data (up to 10,000 points per 1 m2 and above), the size of the segment can decrease to 100 and even 50 meters - with a larger amount, RAM overflow will occur. This is a possible situation when implementing large mobile scanning projects.

3.2 Archiving and spatial roughening

Partially, the problem of a small data segment, as well as the problem of reducing the total amount of stored data, can be solved by combining a decrease in the spatial accuracy of the data and dynamic archiving of information.

The final accuracy of a significant part of airborne laser scanning systems rarely exceeds 20 mm, and mobile scanning systems - 3 mm. However, in most cases, data storage uses a large number of decimal places, up to tenths of a millimeter. Of course, with the above-described real accuracy, this is an overly detailed record of coordinates.

If we estimate the real accuracy and the necessary data density, we user can make certain optimization. For example, to limit the detail of recording coordinates to 1 cm. In addition, filtering is possible. In this example the points at a distance of 0.1-1 cm from each other are considered as one point (if the parameters of surveying do not give reason to expect such density of data). Due to this, it is possible to compress the volume of data by 20-40%. A similar approach is implemented in Bentley Point Tools, and partially can be implemented in TerraSolid and some other software products as well.

The coordinates of the points change quite slowly - for each subsequent point, most of the digits in the coordinates and timestamps coincide with the previous and subsequent points. This feature creates good opportunities for implementing data archiving algorithms. In combination with the previously described approach, such methods allow you to compress the original data cloud 2.2-3.5 times (sometimes even more) in comparison with the original data volume.

A similar approach is used when storing data in LAZ format. Naturally, this causes a decrease in processing speed, but allows to reduce the load on storage systems. And this methods does not solve the problem of non-topological data.

3.3 Geodatabase

One of the most powerful features of modern GIS packages (for example, ArcGIS) is the storage of spatial data in the form of a geodatabase. This approach solves the problem of segmentation and reprojection of materials, including the laser scanning data (also as non-topological points). The problem in this case is extremely slow operation with laser data. The reason is that each point is treated as a separate GIS object, with its inherent attributive characteristics, and processed with all care. The price for this is a very large storage capacity and a slow speed. In practice, working with ArcGIS with large amounts of laser data is extremely inconvenient. There are currently no fundamental ways to accelerate ArcGIS work with geodatabases.

3.4 Quadtree and octree

The quadtree (4-tree, Q-tree) is a data structure in computer science, in which each node has exactly 4 descendants [7]. Quadrant trees can be used to recursively split two-dimensional space into 4 regions. Each area can be a square, a rectangle, or have an arbitrary shape. The term quadtree was introduced in 1974.

At its core, the quadtree allows you to split the space into adaptive cells, depending on the distribution of information. Similarly to a quad tree for two-dimensional space, for three-dimensional space, it is possible to build a similar structure - an octree (divides three-dimensional space into 8 octants).

The ideology of storing and working with data in the form of a quadtree is used in the well-known Google Earth application for hierarchical ordering and addressing of data elements - small raster images that form the final mosaics of satellite images when working in this application.

A partial approach to similar work with data is used in the RiProcess (Riegl) software, and is also used to accelerate the visualization of point clouds in Bentley Point tools, TerraStereo and Riegl Riscan Pro. However, these software environments are unsuitable for storing large infrastructures of spatial by a number of reasons. So, in Riprocess it is very difficult to export a part of the data by a spatial query - for example, exporting data within the specified polygon. There is no way to unlimitedly build a data array without rebuilding a quad tree across the entire data volume. And, most importantly, quad trees in these software products are precisely two-dimensional, and not three-dimensional structures. Nevertheless, this software package has already implemented an approach to calculating statistical characteristics for each cell.

4 Proposed approaches to solving the existing problems of storing laser scanning data

4.1 Common issues

Taking into account the precious issues, we consider it necessary to note that none of the software environments (including modern GIS packages) offers a reasonable solution for storing and working with large amounts of laser scan data in point form.

Previously, such problems took place and were successfully solved in geoinformatics by gradually unifying the forms and formats of data and their storage algorithms (vector graphics, storage in geographic coordinates, layered representation of data, open formats and algorithms). This process took place actively from 1980 to 2010 and successfully lead us to appearance of a large number of GIS packages that make it easy to manage data of any size and spatial coverage, while providing the ability to export data from one program to another without loss of information. A similar step must be taken in LIDAR data storage. Requirements for storing Big LIDAR Data based on 3D points can be formulated as follows:

- the possibility of unlimited increase of data coverage up to global;
- the ability to quickly visualize (in any form) spatial data at any level of datail (up to the consideration of individual points and work with them);
- the possibility of merging various LIDAR data sources without reprocessing whole volume of previously stored information ;
- the possibility of spatial queries and export of individual pieces of LIDAR data;
- well-developed system of visualization of materials in 2D and 3D form.

Obviously, we can conclude that it is the ***form*** of data storage that determines the possibilities for achievement and processing. Indeed, if you look at the history of formats and technologies, it is the flexibility of the format that determines the acceptability of a particular data processing technology. For example, JPG and MP3 formats can be noted. The first opened the way for an explosive increase in the number of stored images and the possibility of transmitting graphics over the Internet. The second revolutionized the field of portable players and online audio streaming. A similar, but less resounding success is in the DXF and GEOTIFF formats.

Creation of the suitable and appropriate data format leads to the rapid development of technologies. The format in this respect is somewhat similar to a cartridge for a firearm. It is the cartridge that is the cornerstone of the rifle complex. Weapons are designed according to cartridge specifications. In the field of spatial data, the situation is similar.

4.2 Estimated specific features of LIDAR data format suitable for Big Data concept.

So, by our opinion, the best way to store Big LIDAR Data for the above purposes is to use an octree. The extent of the dataset is a sphere with the dimensions of the Earth. Working with data stored in geographic coordinates (degrees of latitude and longitude) always encounters the problem of translation into any projection and generates difficulties associated with distortions of projections. Therefore, the most convenient - from the point of view of calculations and the subsequent transition to other coordinate systems - is the Cartesian coordinate system (the origin is in the center of the Earth, coordinates in meters). In this case, the initial space is divided into the initial 8 cubes with a side length of about 8388.608 km (zero level of detail), and then at each level the dimensions of the cube are reduced by 2 times. At the 23rd level, the length of the edge of the cube will reach 1 meter, at the 30th – 1.25 cm, at the 33rd – 1.56 mm.

In the format, it is necessary to establish a certain spatial limit, which can not be exceeded while decreasing the cube size [8]. With a more detailed approximation in 2D mode, the cube is displayed as a pixel and is no longer detailed, when approaching in 3D mode, points inside the cube are displayed (all). When moving away from the minimum possible cube in 3D mode, it is displayed as a single point, with further hierarchical coarsening as it moves away [9]. For a Digital Earth has global coverage, a reasonable cube size can be 0.25-1.0 m, for local datasets it can be 10 cm or even less. We call a cube of a space of similar size an elementary cube.

4.3 Statistics and metadata

For each cube, the statistics of 3D points can be calculated (in advance, when creating the file). This can be the “center of gravity” of the cube (the average position of all its points), the calculation of the average coordinates of the points and their standard deviations for each axis, the averaged color, the amplitude, reflectivity, and the coordinates of the averaging plane (the plane, the sum of the squares of the distances to which from each of the points in the cube is minimal). For each cube, its geographical coordinates are calculated in advance, as well as the coordinates in a widespread projection (UTM). Similarly, the ellipsoidal and orthometric heights of the center of gravity of the cube (adjusted for the geoid model) and the height difference within the cube are calculated.

While visualizing such a data structure at any of its sections in 2D (top view) or 3D mode (isometric view) will look like an array of points, where in each pixel of the screen one cube is displayed by one point. The color of the point is determined by the selected display method (color scale) and can display the height, the spread of heights within a given pixel, the amplitude of the reflected signal, and so on (Fig.2).

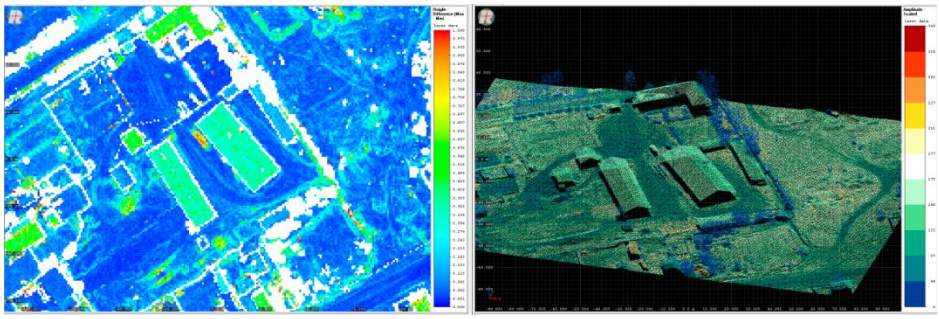

Fig. 2. Riegl RiProcess 2D quadtree data storage. Left – 2D view, quad cells (1 m elementary cells) can be seen. Statistic parameter, used for visualization – height differens of all points inside the cell. Right – the same region visualized in 3D mode by switching to points, stored inside quadtree cells. Colored using attribute «Amplitude»

In addition to spatial properties, metadata is stored for the cube. To store the date and time of receiving the points (if there are multiple shots in the cube, it is displayed in a certain reserved color, and if they are absent, in accordance with the selected

color scale, which does not include the reserved color). The time storage detalisation for the cube should be one day; for individual points, the exact GPS time of point measurement is stored, allowing their additional processing of laser reflection points. It may also be provided for storing textual information about the data source, their accuracy and other similar characteristics, which can be stored in reserved data fields.

4.4 Approaches for saving and merging 3D BIG LIDAR DATA volumes on the drive

Storing data in a format whose ideology is described above will certainly run into the difficulties of storing infinitely large amounts of data in one file. Technically, this is possible. Nevertheless, in Google Earth the similar task (storage of the huge image cache) is solved by storing a large number of small files (pieces of rasters), and when replenishing the system, it is simple to add more detailed pieces of rasters to the folder to existing rasters.

Similar approach can be implemented in this case. Of course, storing data in the form of “one elementary cube - one file” is unreasonable. But it is quite possible to introduce storage of larger cubes (100-1000 m in size) in the form of separate files, and already inside them to save data of elementary cubes. We call such a larger cube a “group cube”.

In this case, when adding a new dataset to the old one, it will be necessary to simply rewrite some new group cubes to a new location, and re-calculate some of them by updating the statistics and metadata of each updated group cube. However, the rest of the data will not have to be recalculated. Exporting of the data can be done in the similar way.

5 Conclusion

Creating a unified data format based on algorithms for working with the octree will allow solving a whole group of BigData problems in the field of laser scanning for Digital Earth: the need to structure large segments of data, quick access to arbitrary sections of data, and also minimize the cost of merging and combining overlapping sections of simultaneous data.

An important feature of the data structure in the form of an octree when storing laser points is the ability to create topological data structures that, in addition to statistical and spatial information, contain information about neighboring data segments.

Undoubtedly, such an organization of information is significantly more complicated than the currently accepted methods of storage in open formats (LAS, BIN, FBI, XYZ). However, without solving this problem, further development and mass implementation of the use and analysis of laser scanning data will be extremely difficult.

References

1. RIEGL VZ6000 homepage, http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/33/ Last accessed 10 Jul 2020
2. RIEGL 1560ii homepage, http://www.riegl.com/nc/products/airborne-scanning/produktdetail/product/scanner/68/. Last accessed 13 Jul 2020
3. RIEGL VMX RAIL homepage, http://www.riegl.com/nc/products/mobile-scanning/produktdetail/product/scanner/67/ Last accessed 19 May 2020
4. Giuliani, G., Chatenoux B., De Bona A.: Building an Earth Observation Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). Big Earth Data 1-2 (1), 100–117 (2017)
5. Chen Q.: Airborne lidar data processing and information extraction. Photogrammetric Engineering & Remote Sensing 2 (73), 109–112 (2007)
6. Rylskiy I.A., Markova O.I., Eremchenko E.N., Panin A.N.: Building object-based virtual models based upon terrain laser scanning and uav data. International scientific review 6 (71), 75–83 (2020)
7. Octree homepage, https://en.wikipedia.org/wiki/Octree Last accessed 16 Jan 2020
8. Anh-Vu V., Truong-Hong L., Laefer D., Bertolotto M.: Octree-based region growing for point cloud segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 104, 88–100 (2015)
9. Wang M., Tseng Y: Incremental segmentation of lidar point clouds with an octree-structured voxel space. The Photogrammetric record 26 (133), 32–57 (2011)