How to Properly Plan the Reduction in the LiDAR Big Dataset?

Wioleta BŁASZCZAK-BĄK and Anna SOBIERAJ-ŻŁOBIŃSKA, Poland

Key words: reduction, big dataset, OptD method

SUMMARY

There are many methods of data collection, which leads to a big dataset (e.g.: LiDAR, bathymetry measurements). Such datasets are difficult or sometimes impossible to rational use. Therefore, in the stage of pre-processing the big dataset is reduced without losing data necessary for the proper implementation of objective study. The process of reducing the big dataset will allow efficient, less time consuming and labor intensive processing. Depending on the purpose of data processing and project requirements the reduction of big dataset must be properly planned. It involves selecting the appropriate method of reducing big dataset, choosing the appropriate tools, criteria and parameters.

The paper presents the stages of proper planning to reduce the size of the set derived from LiDAR. It also presents an original method for reducing called the Optimum Dataset.

SUMMARY in Polish

Istnieją różne metody pozyskiwania dużych zbiorów danych (np. LiDAR, pomiary batymetryczne). Takie zbiory mogą być trudne lub czasem wręcz niemożliwe do racjonalnego wykorzystania. Dlatego w etapie przetwarzania wstępnego duże zbiory danych mogą być redukowane, ale w taki sposób aby nie utracić danych niezbędnych do prawidłowej realizacji postawionego celu opracowania. Proces redukcji dużych zbiorów danych pozwala na efektywne i mniej pracochłonne przetwarzanie. W zależności od celu opracowania i założeń projektowych redukcja powinna być prawidłowo zaplanowana. W szczególności powinna zostać wybrana odpowiednia metoda redukcji, odpowiednie narzędzia, kryteria i parametry. W pracy przedstawiono etapy prawidłowego planowania redukcji liczebności dużego zbioru danych pozyskanego z LiDAR. Zaprezentowano również nowatorską metodę redukcji danych Optimum Dataset Method.

How to Properly Plan the Reduction in the LiDAR Big Dataset?

Wioleta BŁASZCZAK-BĄK and Anna SOBIERAJ-ŻŁOBIŃSKA, Poland

1. INTRODUCTION

Data reduction is a procedure to decrease the size of the dataset in order to make their analysis more effective and easier. There are several ways to conduct a reduction. It may involve, among others, the use of advanced statistical methods that allow to reduce the large dataset to fundamental factors, dimensions, or clustering, explaining the important relationship between the analyzed variables/values.

Another way of reducing the dataset is conversion of a large number of variables into one value common to all. Reduction can be also conducted by removing a predetermined number of data, but in such way, that remaining reduced set is representative for given population (set).

During reducing, user should be aware of the development and the quality of the product which will be built on the basis of the reduced dataset. Due to these aspects two methods for decreasing the set can be used: generating or reduction. Generation involves creating a grid of regular figures. The grid nodes are interpolated, they have new coordinate values. (Gościewski 2013, Bauer-Marschallinger et al. 2014). These coordinates are calculated on the basis of measurement data located in vicinity of interpolated points. Reduction decreases the size of dataset by removing some points according to given algorithm, remaining points are original points from measurement (Błaszczak 2006, Błaszczak et al. 2011a, 2011b, Chen 2012).

Generation is very often used to build the DTM (Digital Terrain Model) or DSM (Digital Surface Model). They are the most popular products of LiDAR data (Light Detection and Ranging) development. These models are created by coordinate interpolation in a regular grid. However it is better approach to reduce the dataset, in order to operate on real (not interpolated) data. Using generation to decrease the set cause the double generation of coordinate, which has an impact on the quality of the created models and studies.
Reduction of the set is an issue that requires proper planning, so the set after reduction meets all the user's expectations.

The scheme of the process of selecting satisfactory or optimal solution presents Figure 1.

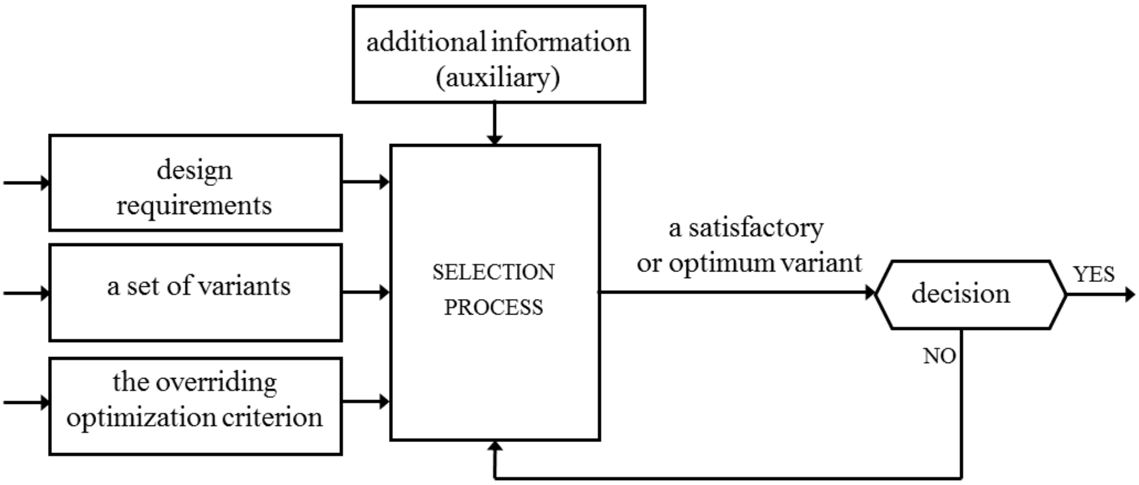
additional information (auxiliary)
design requirements
a set of variants
the overriding optimization criterion
SELECTION PROCESS
a satisfactory or optimum variant
decision
YES
NO

Figure 1. The scheme of choice process (source: Ostwald 2005)

Of course it is better if the result is an optimal solution in terms of adopted criteria. Within reduction methods, which provide the optimal solution there is the Optimum Dataset method (OptD) proposed by Błaszczak-Bak (2016). The OptD method allows for obtaining a representative sample of the original dataset as an optimal set of LiDAR,. The proposed method put an emphasis on fact, that the reduction of large datasets is conducted in a way, that the information necessary for the proper performance of a task is not lost. Application of the OptD method in preparation of the data for DTM construction was more accurate and less time-consuming. It allows for the effective DTM generation and reducing the time and cost of LiDAR point cloud processing, what in turn enables to conduct efficient analyses of acquired information resource.

2. OPTIMUM DATASET METHOD

LiDAR point cloud development based on OptD method in the form of flow chart is presented in Figure 2.

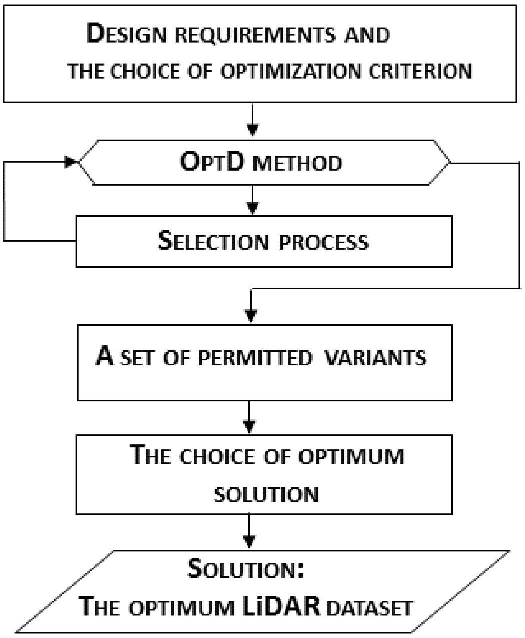
Design requirements and
the choice of optimization criterion
OptD method
Selection process
A set of permitted variants
The choice of optimum solution
Solution:
The optimum LiDAR dataset

Figure 2. LiDAR point cloud development based on OptD method (source: own study)

OptD method can be conducted in two variants:
(1) OptD method with single objective optimization called OptD-single,
(2) OptD method with multi objective optimization called OptD-multi.
If OptD-single method is chosen, then a set which strictly fulfilling one condition is sought. If there is a decision on processing using OptD-multi, then in result several sets will be obtained, among which the best one should be selected.
The algorithm of OptD-single method consists of the 12 steps, while the algorithm of OptD-multi consists of the 16 steps. All steps are presented below:
step 1: Loading the N points of the original LiDAR dataset.
step 2: Establishing optimization criterion (f), e.g.: the number of points in the set, the mean error of DTM. For OptD-single we choose 1 criterion, for OptD-multi we choose minimum 2 criteria.
step 3: Determination of the XYZ coordinate system. The aim of this step is to 'fitting' the measurement lines, that are a direct result of the measurement, into coordinate system so that the measurement lines are approximately parallel to the X or Y axis (the coordinates of points have a certain regularity and are arranged in a measurement lines).
step 4: Projection of LiDAR data points onto a plane X0Y.
step 5: The choice of initial width of strips (L). Choosing the appropriate strips width some parameters (depends on the user) can be taken into consideration: the average distance between points in the measurement set, as well as the distance between the belts, which arose directly from the type of measurement (here: LiDAR) and they are a consequence of the characteristics of LiDAR measurement (height, velocity, scan angle etc.). Another way of strip's width determination is in an iterative process and change e.g. at a fixed interval.
step 6: The division of area covered by points on the test strips (nL).
step 7: Selection of measurement points for each measurement strip.

step 8: Projection on the plane Y0Z LiDAR data points for each measurement strip.
step 9: Selecting the method of cartographic line generalization, e.g. the Douglas-Peucker (D-P) (Douglas, Peucker 1973) or Visvalingham-Whyatt (Wisvalingham, Whyatt 1992).
step 10: Using the selected method of generalization in the plane Y0Z. Choosing the tolerance parameters in the selected method of generalization. For the method of D-P it is a distance of tolerance. The initial value of the section is defined by the user, the following values are determined in an iterative process, in which there is increase or decrease at a fixed interval.
step 11: Obtaining the reduced data set with the number of M, where M<N.
step 12: Verification, whether obtained in Step 10 set fits the specified criterion optimization. If so, the reduction process is completed, and the obtained set from Step 10 is the optimal dataset. If not, the steps 9-12 are repeated, wherein in step 8 the value of tolerance parameter is changed. If repeating steps 9-12 do not give a solution, there is need to back to Step 6 and change the width of measuring strip. STOP algorithm for OptD-single. The next steps are used only for OptD-multi method.
step 13: Save the reduced set to the solution set.
step 14: The decision whether to process the set again in order to obtain a further reduced set of data points. If the decision is positive another reduced dataset will be added to the solution. If not, the next step of the OptD-multi method is performed.
step 15: In the solution set, the gathered (reduced) data sets are checked. The condition of Pareto is checked.
step 16: Selecting the optimal solution in the sense of Pareto. The resulting set meets the given optimization criteria. It contains fewer points and certainly is suitable for generating DTM.
In order to properly carry out the reduction one should takes into account the principles of optimum design, such as (Oswald, 2005):

- the possibility to choose the solution (variant, decision) from a set of acceptable solutions.
- designer has created by himself or imposed from above, conditions allowing for the evaluation of solutions (decision).
- designer can decide, at least for some of the solutions, which could be considered to be better; which he favors due to the adopted conditions.
- designer can justify and defend his choice, which, in his opinion, is optimal.

The proposed OptD method complies with the principles of optimum design. Step 14 of the algorithm makes it possible to decide whether we need another reduced set in the set of acceptable solutions.
Step 15 and step 16, in turn, gives the opportunity to choose the perfect set from the point of view of the designer or to select a solution in the sense of Pareto.

Another important issue during planning a reduction is the decision of when to use a reducing method.
It also applies to the OptD method proposed mainly to reduce the LIDAR data. Planning to use of the OptD method for data from airborne laser scanning (ALS) for generating a DTM can be carried out in two variants, which shows the scheme in the Figure 3.

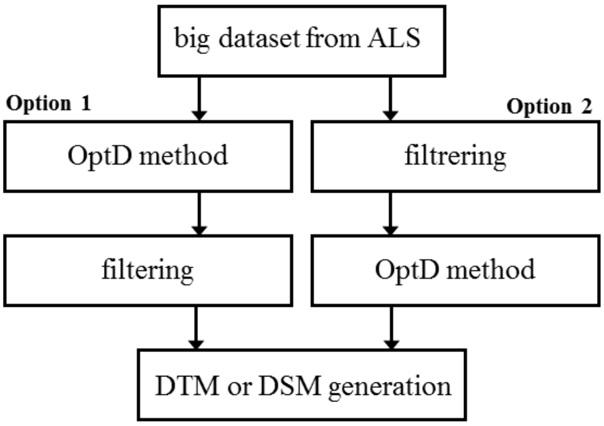

Figure 3. The order in the processing of big dataset from ALS (source: own study)

3. MATERIALS OF THE RESEARCH

The study area is a section of the national road No. 16, Sielska Street in Olsztyn, located in Warmia-Mazury. Airborne laser scanning was made by Visimind Ltd. Fragment of this measurement was selected for tests. Laser scanning angle was 60 degrees, with a frequency of 10,000 Hz scanning. Scanning was performed from a helicopter with speed of 50 km/h at an altitude of 70 m. Selected test area was called **Object** and it contains 12781 points.

Original point clouds and point clouds after filtration, which will be used to generate the DTM are shown in Figure 4.

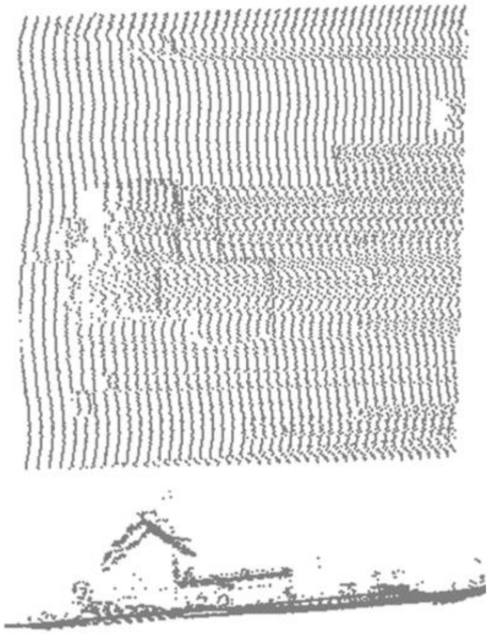

Figure 4. Test area Object (original dataset)
(source: own study in CloudCompare v.2.6.0)

3.1 Option 1

In option 1 "filtering – OptD method" (F – O) the selected section was filtered by using adaptive TIN model method (Axelsson 2000) in own software. As a result of the filtration, there are two sets of data for Object: the set of points showing the topography (topographic surface dataset - TSset) and a set of points showing the detail points.

The topographic surface dataset for Object was called **TSset1**. The number of points in this set is 10414. The application of the OptD-single method resulted in obtaining the optimum solution, which contained **OptDset1** with 8121 number of points (22% of points were removed). The **TSset** after OptD-single method is presented in Figure 5.

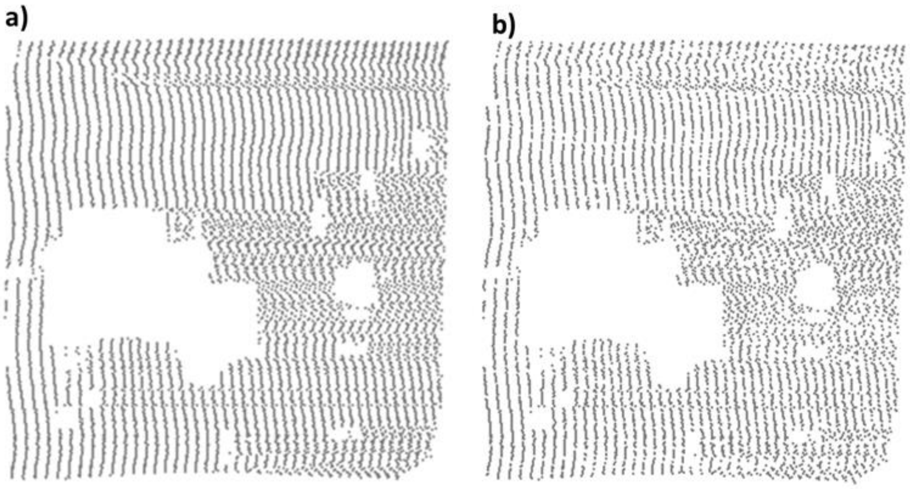

Figure 5. a) dataset after filtration (**TSset1)**
b) **TSset1** after the OptD-single method application **(OptDset1)**
(source: own study in CloudCompare v.2.6.0)

3.2. Option 2

In option 2 "OptD method – filtering" (O – F) the selected section was optimizated by using OptD-single method in own software and the next step was filtering by using adaptive TIN model method. As a results of the OptD-single method there is a one set of data which is called **OptDset2**. The number of points in this set is 9808.

The application of adaptive TIN model selected the dataset with points which represented the topography: **TSset2**. The TSset2 consists of 8005 points.

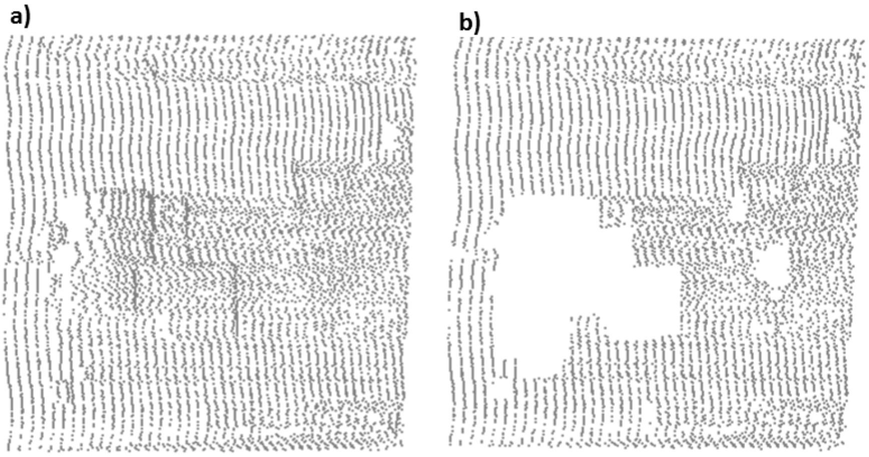

Figure 6. a) dataset after OptD-single method application **(OptDset2)**
b) **OptDset2** after filtration (**TSset2**)
(source: own study in CloudCompare v.2.6.0)

3.3. Analyzes

For **TSset1** and **OptDset1** and **TSset2** and **OptDset2** the parameters were calculated:

- mean error (m_0):

$$m_0 = \sqrt{\frac{\sum(z_{mean}-z_i)^2}{M-1}} \qquad (1)$$

where: z_{mean} is a mean height calculated from heights of both DTMs, z_i (i=1,2…, M) are heights of the point assumed for creating DTM, M is the size of the set used for DTM construction,

- range (R):

$$R = z_{max} - z_{min}, \qquad (2)$$

where z_{max} is the maximum height and z_{min} is the minimum height.
Results of processing ALS point cloud based on OptD-single method are presented in Table 1.

Table 1: Results of processing with the OptD-single method

Parameters:	**Option 1**	**Option 2**
Total number of points in Objects	**12781**	**12781**
Number of terrain points in **TSset**	**10414**	**8005**
Number of terrain points in **OptDset**	**8121**	**9808**
Total operation time [sec.]	**358**	**250**
m_0 **TSset** [m]	**0.648**	**0.701**
m_0 **OptDset** [m]	**0.654**	**0.698**
m_0 **TSset** – m_0 **OptDset** [m]	**-0.006**	**0.003**

z_{mean} in **TSset** [m]	**132.105**	**132.081**
z_{mean} in **OptDset** [m]	**132.109**	**132.062**
R **TSset** [m]	**3.429**	**2.819**
R **OptDset** [m]	**3.430**	**19.980**

On the basis of the dataset obtained from option 1 and option 2 DTMs were generated. It is presented in Figure 7 (F - O) and in Figure 8 (O - F). For this models grid with 1m size was adopted.

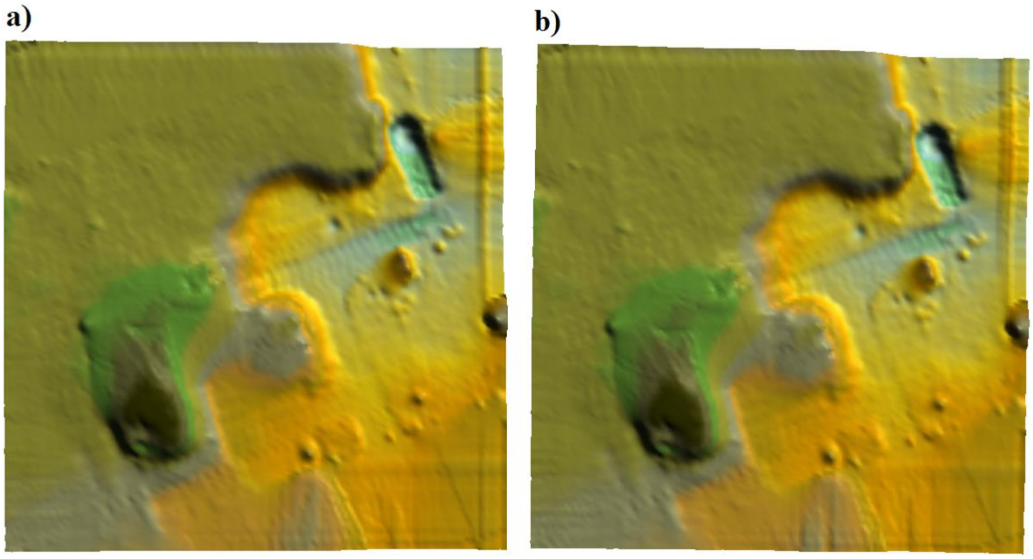

Figure 7. DTMs a) from all points of **TSset1**, b) from points of **OptDset1** (source: own study in Surfer v.8 Demo)

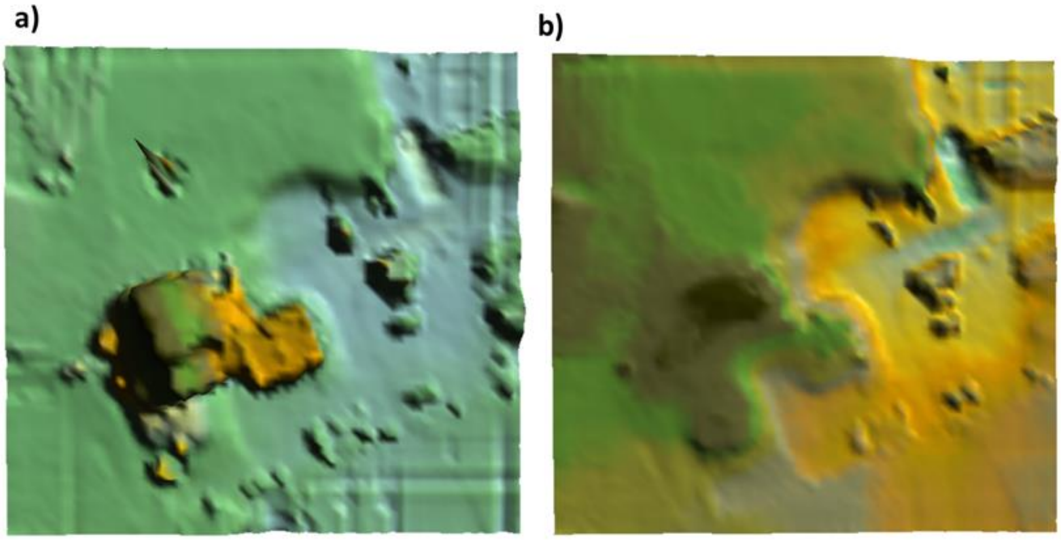

Figure 8. DTMs, a) from all points of **OptDsetD2**, b) from points of **TSset2** (source: own study in Surfer v.8 Demo)

Finally, the result of option 1 is a DTM generated from a set of OptDset1, the result of option 2 is DTM generated from a set of TSset2. For DTMs, called as DTM$_{F-O}$ and DTM$_{O-F}$ respectively, accuracy analyses were performed. For models generated for DTMs quality parameters were calculated:

- mean value of height difference (Δh_{mean}):

$$\Delta h_{mean} = \frac{\sum_{i=1}^{M}\left(Z_{DTM_F} - Z_{DTM_{O-F\ or\ F-O}}\right)}{M} \tag{3}$$

where:

$Z_{DTM_{TSset}}$ – heights of nodes generated from original numbers of points,

$Z_{DTM_{OptDset}}$ – heights of nodes generated from reduced numbers of points,

M– the size of the set used for DTM construction.

- root-mean-square error (RMSE), which describes the absolute altitude accuracy of DTM:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\left(Z_{DTM_F} - Z_{DTM_{O-F\ or\ F-O}} - \Delta h_{mean}\right)^2}{M}} \tag{4}$$

- coefficient of determination, which is the measure of model adjustment (the closer to 1, the better the match of the model to another model):

$$D^2 = \frac{\sum_{i=1}^{M}\left(z_{DTM_{O-F\ or\ F-O}} - z_{mean}\right)^2}{\sum_{i=1}^{k}\left(z_{DTM_F} - z_{mean}\right)^2} \tag{5}$$

where: z_{mean} is a mean height calculated from heights DTM$_F$.

- mean error (M_0):

$$M_0 = \sqrt{\frac{\sum(Z_{mean} - Z_{DTM})^2}{M-1}} \tag{6}$$

where: Z_{DTM} are heights of the point of DTM$_{O-F}$ or DTM$_{F-O}$.

The obtained results are presented in Table 2.

Parameters:	**DTM$_F$**	**DTM$_{F-O}$**	**DTM$_{O-F}$**
Δh_{mean}	0.001	-0.001	0.0005
RMSE	0.014	0.015	0.026
M_0	0.646	0.649	0.725
D^2	-	0.999	0.992

The analysis shows that the DTM$_{F-O}$ is much more fitted to the DTM generated on the basis of the only filtered set, although in this set number of points is slightly higher. (8121 points in the set for generation in option 1, 8005 points in the set for generation in option 2). RMSE and M_0 are better (smaller) for DTM$_{F-O}$.

4. CONCLUSIONS

This paper presents the stages of proper planning to reduce the dataset derived from LiDAR. It also presents an original method for reducing called the Optimum Dataset. The algorithm of this method can reduce the dataset in terms of number of measuring points for a given optimization criterion. In this paper two options were tested: filtering – OptD method, OptD method – filtering. This two options can be used for reducing.

Based on the analysis, following general conclusions can be stated:

1. Innovative the OptD method is a simple in application method for data reduction, which takes into account optimization criteria.
2. The result of the implementation of the OptD method is an optimal dataset that can be used to generate DTM.
3. The OptD method fulfills all the expectations of reducing the size of the dataset without losing data necessary for the proper DTM generation.
4. The algorithm of new method implies, that only once the initial values of the parameters are introduced, the subsequence values of them in following iterations are automatically selected.
5. Option 1 gives the better solution then option 2.

REFERENCES

1. Bauer-Marschallinger B., Sabel D., Wagner W.: 2014, Optimisation of global grids for high-resolution remote sensing data. Computers & Geosciences, 72 (2014), 84 - 93. DOI: 10.1016/j.cageo.2014.07.005
2. Błaszczak, W.: 2006, Optimization of large measurement results sets for building data base of spacial information system. Doctors thesis, University of Warmia and Mazury in Olsztyn.
3. Błaszczak-Bąk W, Janowski A., Kamiński W., Rapiński J.: 2011a, Optimization algorithm and filtration using the adaptive TIN model at the stage of initial processing of the ALS point cloud. Canadian Journal of Remote Sensing., No. 37(6), pp. 583-589. DOI: 10.5589/m12-001
4. Błaszczak-Bąk W, Janowski A., Kamiński W., Rapiński J.: 2011b, ALS Data Filtration with Fuzzy Logic. Journal of Indian Society of Remote Sensing, 39, 2011, 591-597. DOI: 10.1007/s12524-011-0130-2
5. Chen Y.: 2012, High performance computing for massive LiDAR data processing with optimized GPU parallel programming.
6. Gościewski D.:2013, Selection of interpolation parameters depending on the location of measurement points. Giscience & Remote Sensing, 50(5), 515-526. DOI: 10.1080/15481603.2013.827369
7. Ostwald M., 2005: Podstawy optymalizacji konstrukcji, Wyd. Politechniki Poznańskiej.

BIOGRAPHICAL NOTES

CONTACTS

Wioleta Błaszczak-Bąk
Institute of Geodesy, University of Warmia and Mazury in Olsztyn
Oczapowski St. 1
Olsztyn
POLAND

Tel.: +48 89 523 33 05
wioleta.blaszczak@uwm.edu.pl

Anna Sobieraj-Żłobińska
Department of Geodesy, Gdansk University of Technology
Narutowicza St. 11/12
Gdansk
POLAND
Tel.: +48 58 347 22 12
annsobie@pg.gda.pl