The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W3-2023
2nd GEOBENCH Workshop on Evaluation and BENCHmarking of Sensors, Systems and GEOspatial Data
in Photogrammetry and Remote Sensing, 23–24 October 2023, Krakow, Poland

THE UTILIZATION OF SYNTHETIC AND SEMISYNTHETIC POINT CLOUDS AND IMAGES FOR TESTING NOVEL APPROACHES FOR CORRECTING LIDAR DATA

K. Pargieła, A. Rzonca*, M. Twardowski

AGH University of Krakow, Cracow, Poland, Faculty of Geo-Data Science, Geodesy, and Environmental Engineering – (pargiela, arz, misiekt)@agh.edu.pl

KEY WORDS: LIDAR, Benchmark, Synthetic Data, Semisynthetic Data, Point Clouds.

ABSTRACT:

The paper presents the application of lidar data and photo datasets, external orientation parameters (EOPs), ground control points (GCPs), and check points for testing new methods of geometric lidar data correction. These datasets are utilized to validate novel approaches such as altimetric deformation methods based on stereo models or lidargrammetric methods that utilize image matching and specialized lidar data formats. The paper presents specific use cases of these data as examples of two tested processes. After describing these processes, the methods of synthetic and semisynthetic data simulation are presented. The simulation is directed and subordinated to the aspects of the new method being tested. The data must be used for testing starting from basic functionality up to specific and untypical cases of new method application. By presenting specific cases of the application of synthetic and semisynthetic data, the paper introduces the general idea of benchmarking based on synthetic and semisynthetic data as another means of validating new methods. These artificially generated datasets provide a controlled environment for evaluating the effectiveness of new methods to be investigated.

1. INTRODUCTION

The process of testing of novel approaches used to be done using real data benchmarks (Du et al., 2015, Mitishita et al., 2020, Parmehr et al., 2013). A statistical selection of different case studies is an advantage of using such a kind of data. We can compare several methods using one benchmark data or we can use several case studies to know more about the method being researched and to test it within this approach. Many different aspects of the method are tested at the same time. The disadvantage is that one has to test many variants of many different cases. For example, 4 variables of 5 specific settings each give 1024 repetitions to process and afterwards to analyse and to interpret the results.

In this paper we present another known approach with some modifications: the simulation of synthetic and semisynthetic data as an integrated datasets of lidar and photogrammetric data. The application of such a synthetic and semisynthetic benchmarks reduces the number of variables and it can serve to compare the methods as well.

Another advantage of this way is that synthetic and semisynthetic benchmark data seems to be used easier for automatic testing than real data, because the results are more evident according to the testing data preparation (Rzonca and Twardowski, 2022, Wang et al., 2019).

2. SYNTHETIC AND SEMISYNTHETIC DATA IDEA

Synthetic data is data generated by a computer or human, it is not captured by any sensor. This kind of data has some general, specific characteristics: the values and order of the data are usually arbitral, geometrically and radiometrically regular, potentially randomized. Additionally, the data can be simplified as much as necessary. It can be prepared for testing of one parameter or one single process. Then it checks the correctness of the process whose result is known a priori and expected. Moreover, the synthetic data can be enriched by real data during the process of its generation: for instance, real orthoimage can be applied for colorizing synthetic point clouds - by adding real-like colours to the points. Another case: 3D point clouds can be centrally projected as a frame image to get lidargramms (Rodríguez-Cielos et al., 2017, Valbuena et al., 2011). Regardless of the origin data, the point cloud generated automatically and colorized by an image of real objects (orthoimage) is a lidar synthetic data - it is an RGB point cloud and vector data, not a raster being only a source of colours. Analogically, the frame image generated as a projection of a real point cloud is an image synthetic data - it is a raster, neither point cloud nor vector.

Semisynthetic data is a simplified version of real data. The simplification changes the real data to testing data easy to apply and to analyse (Schofield et al., 2022). The semisynthetic data keeps the same character: lidar point cloud is still a point cloud and an image stays an image. The only difference is that the semi synthetic data gives strict and evident results during testing of specific method.

Based on above rules the benchmarks of synthetic and semisynthetic data can be predefined and prepared.

3. SYNTHETIC BENCHMARK DATA

3.1 Preparation

There are five basic phases of benchmark synthetic data preparation for further use: definition of the characteristics of the data, process of generation, transformation (or disruption), evaluation and publication.

* Corresponding author

3.1.1 Phase 1: Data definition: Synthetic data are defined by the process and/or the parameter to be tested (table 1). Firstly, one decides whether the data are a block of images, point cloud (in strip of block), other vector data and additional data like ground control points (GCPs) or metadata like external orientation parameters (EOP), internal orientation parameters (IOP) or the trajectory of the sensor platform. There are three options of preliminary decision of the geometry and radiometry of the data. It can be regular, randomized or real-like. For instance, for testing of lidar altimetric correction processes the synthetic data can be a rectangular point cloud with equal height and ground control points located in a regular pattern. We can control the process strictly using such data and the test result can be binary.

3.1.2 Phase 2: Generation: The synthetic images present a regular pattern or texture. They can be generated from scratch or they can be based on some converting process of different data. The image without any real-like colours can be generated by the script using predefined parameters like ground sample distance (GSD) and IOP of virtual camera (pixel size, focal length and image size). Based on these parameters the synthetic data and metadata are generated. The real-colour images can be acquired from lidar data, so named lidargramms. According to the initial data, RGB or intensity values of pixels are interpolated and projected to an image matrix using a collinearity equation.

The synthetic point cloud can be generated as a regular grid with one Z value in the simplest case. The XYZ coordinates can be also randomized. The point cloud can be generated with mono colour or natural, interpolated colours. It is necessary to define parameters for this data generation: coordinates XY of origin of the point cloud, its density and height (for mono colour points), additionally IOPs and EOPs of photograms (for real RGB acquired from image), and XY of origin and GSD for RGB acquired from ortho.

	Kind of data	Colour	Parameters	Add. data Metadata
1.1.	Image	Pattern	GSD, pixel size, focal length, image size	EOPs, GCPs
1.2.	Image	From lidar	GSD, pixel size, focal length, interpolation mtd, image size	EOPs, GCPs
2.1.	Point cloud	Unique RGB or intensity value	Origin, Density, height	GCPs
2.2.	Point cloud	RGB from ortho	Origin, Density, height, ortho	GCPs
2.3.	Point cloud	From images	Origin, density, height, image, IOPs, EOPs	GCPs

Table 1. Synthetic data.

3.1.3 Phase 3: Data transformation and controlled disruption: This phase includes the final preparation. According to the specific needs, the data can be additionally transformed. This transformation should be understood as geometric change but also radiometric change as well. There are three basic possibilities of this phase: to omit transformation, to transform the data using the same tested process in the opposite direction or to transform the data by another process.

The first option: the a priori generated data is ready to use for testing. In this case the process of generation includes the specification of the process to be tested.

The second option has two stages: the first is deformation, and the second is back correction. It is processed by the same algorithm in the opposite direction. After both stages one should get the same data as the data at the beginning, in some cases in limits of the accuracy of the calculations.

The third option is to deform data using another process to get the data of expected values to be changed back by the process being tested. This method seems to be the most objective because the algorithms of deformation and correction are different and independent.

3.1.4 Phase 4: Evaluation and Phase 5: Publication: The next phase is a phase of control.

Firstly the format of the synthetic data has to be appropriate to the further process.

Secondly one has to compare the values of the data after generation and after transformation. It should be compared also with the control data.

Next check's aspect is a parameters check. Before testing we should be sure that all variants are correctly prepared using appropriate values of parameters and these parameters were applied. After this one should be sure that all variants represent the full spectrum of values to be tested.

The last part of this phase is a completion check. All necessary data for testing should be complete and ready for further processes ordered in an easy for use way during testing and without unnecessary redundancy.

The last phase is publication of the benchmark data. It can be done with any repository or data storage as google disk or github.

3.2 Examples of benchmark synthetic data

The simulated lidar data can be used for evaluation or analysis of several lidar processes. For example we present data created and applied for lidar data altimetric and 3D flexible deformation. Firstly we present some options for lidar data simulation, later for photo and auxiliary data generation.

3.2.1 Lidar data:

1. One-color point cloud

For altimetric deformation processes one can use data without any additional radiometric information. The pure geometry is enough for this purpose. The simplest example for simulated lidar data is a rectangular, flat set of points (fig. 1). The definition includes equal height and colour of the points, coordinates of the origin and density or X- and Y-resolution.

After generation the coordinates of the cloud are checked and a separate file of GCPs is generated. An additional file of checkpoints (CHPs) is generated as well. When GCPs are reference data and they define the demanded corrections of the height, the CHPs are used for correctness checks of the results in space between the GCPs. So GCPs application is a data transformation phase, and CHPs are used for evaluation of lidar altimetric deformation processes in height.

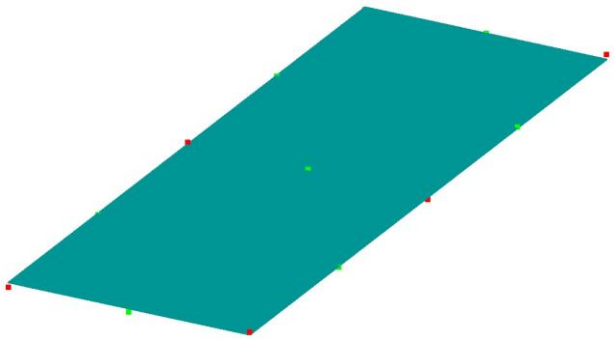

Figure 1. Rectangular flat point cloud: red dots - GCPs and green dots - CHPs.

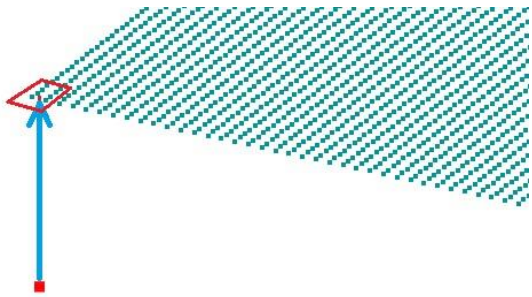

Figure 2. Simulated difference between GCP and flat point cloud patch.

2. RGB point cloud

RGB point cloud is simulated for evaluation of processes of three dimensional lidar data correction. The generated point cloud in strips is colorized by orthoimage. The real-like colours are used to localize points also in the XY plane, not only in the Z direction as it was possible before.

Figure 3. Two strips of RGB point clouds with GCPs (red dots).

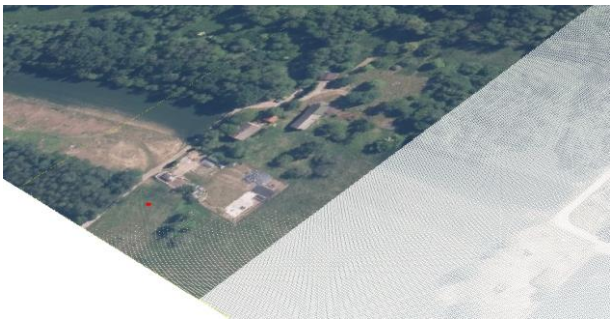

Figure 4. Zoomed strip's overlap area with GCP (red dot).

Further processing of the point cloud like lidargrams' generation and their matching is possible thanks to the color applied from the orthoimage.

3.2.2 Photo data:

1. Single image

Single images are used for colorizing synthetic point clouds to add the radiometric information to the pure geometry. The colour is necessary for localizing the XYZ GCPs and CHPs. The important parameters during image simulations are GSD and image proportions. The GSD should be selected with attention to the lidar data density.

The images can be generated like a camera calibration test field, black and white chessboard) (fig. 5), or RGB chessboard (fig. 6). RGB chessboard is much better for specific GCPs and CHPs locations.

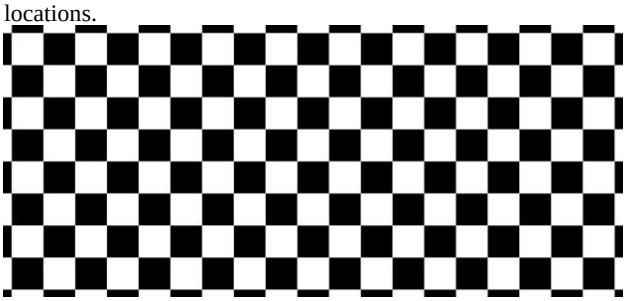

Figure 5. Black and white chessboard.

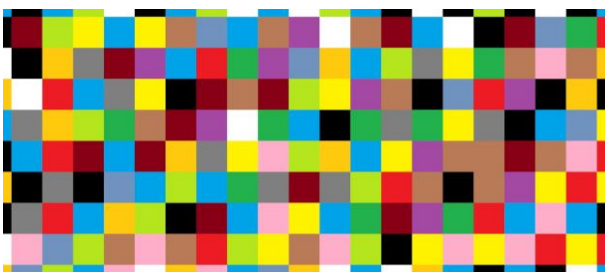

Figure 6. RGB chessboard.

Natural colour of the terrain can be applied from orthoimages to colorize flat synthetic point clouds. The full section of orthoimage can be cut to strips identical in shape to the synthetic lidar data strips (fig. 7).

Figure 7. Selected orthoimages for lidar data colorizing.

2. Block of lidargrams

Lidargram generation is a next potential option of synthetic photo data generation. Lidargrams are synthetic images of point clouds virtually captured by camera of arbitrary IOPs and EOPs, that can be predefined according to the specific needs. Lidargrams can be used in processes of lidar data enhancement and they can be matched (also densely matched) and aerially triangulated as normal aerial imagery. Parameters of their generation are subordinated to the quality of the process of image matching.

Figure 8. Strip of lidargrams of 60% overlap.

3. Photogrammetric metadata: IOPs, EOPs

There are two additional datasets for blocks of lidargrams to enable a use of these lidargrams analogically to real photogrammetric data.

The first is a set of interior orientation parameters of a virtual camera: focal length, pixel size and frame size. The focal length is set to optimize the geometry of the photogrammetric intersection - the angle between the homologous rays. The pixel size and format of the frame is defined to present all the synthetic points of the lidar data on the image without any loss of information. Generally these parameters are used to generate the lidargrams with optimal quality. The optimization of lidargram generation is a separate future research problem.

External orientation parameters are defined to project the lidar data on the frames correctly. They are calculated starting from the input data like: virtual flight height over the terrain, strip width, omega and phi angles equal zero and central line of this flight defined by start and end XY points. The coordinates of virtual centre of projection (EOPs) are calculated taking into account also the overlap of lidargrams.

3.2.3 GCPs: Ground control points (GCPs) are necessary for lidar and photo synthetic data. GCPs as terrain points with global coordinates are compared with their representation on the lidar or photo data. The discrepancies present the values of corrections to apply for data quality enhancement. The possibility and accuracy of measurement of GCPs location on lidar data or lidargrams depends on two main factors: the definition of gcps and quality of the point cloud (density, colour) or lidargrams (resolution, pixel size).

3.2.4 Trajectory: Trajectory is data that can be simulated as an integral lidar data or as a standalone data for different purposes.

Coordinates of aerial lidar points are a function of trajectory and measurements done in specific, registered as timestamp time points. There are two options for transforming lidar data. The first is to change the coordinates of the points (in a rigid or nonrigid way), the second - to change the basic observations like trajectory data. The first is possible thanks to the methods based on theory of altimetric deformation of stereoscopic models or using lidargrams, the second is used in most software for lidar data processing. Trajectory is a list of GPS and INS observations with timestamps. Distance and angle are measured for each point and based on a specific point of trajectory. The point is defined by 6 EOPs as the observations. One can transform the lidar strip correcting these observations. It is possible to simulate the point cloud starting from points and trajectory and calculate a posteriori angles and distances. The trajectory can be randomized before.

Another way of trajectory simulation is a generation of standalone trajectory of the platform of different sensors like RGB, multi- and hyperspectral cameras and scanners. In this case the trajectory can be generated as a base data, and all data of the other sensors should be generated taking into account simulated sensors' offsets and the trajectory.

3.2.5 Idea of integrated synthetic benchmark datasets: The idea of benchmarking by synthetic data of many sensors comes from the idea of simulation of trajectory and further all data captured by the sensors mounted on one platform or regarding one virtual test area.

4. SEMISYNTHETIC BENCHMARK DATA

4.1 Preparation of the data

Analogically to the synthetic data, there are 4 phases of such a data preparation: definition, selection and completion, transformation (or disruption), testing.

4.1.1 Phase 1: Data definition: The testing based on semisynthetic data requires a simplification of the real data, but in some aspects the data have more parameters of random propagation. This kind of data has to be defined for testing specific parameters and parallelly they have to be similar or identical to real data in other aspects. These aspects make alike the data and their processing to real data processing.

4.1.2 Phase 2: Selection and completion: Real data is a source of semisynthetic data. The representative part of the real dataset has to be selected. For lidar data the parts of strips can be selected. In case of using the block data, overlapping strips are selected. The only difference with real data until this stage is a specific selection or simulation of lidar strips by rectangular parts of the lidar block.

4.1.3 Phase 3: Data transformation and controlled disruption: This phase includes the final stage of preparation. According to the specific need, the data can be additionally transformed. There are three basic possibilities of final data preparation for specific process testing.

The first possibility is to omit the transformation of the point cloud and to simulate its deformation by arbitrarily defined discrepancies by GCPs and CHPs coordinates. The same can be done for block of images without deformation of EOPs. The effect can be managed by EOPs' or IOPs' changes.

The second possibility is to transform the data using the same correction process that will be used in the opposite direction. To deform and to correct back. This option might permit checking the process effectiveness and its accuracy after back processing. The data after both direction processing should be the same.

The third last option is to use another deformation process to change the data and correct them by testing a new method. The advantage is that an independent process is applied for deformation and this option simulates a more general case.

4.1.4 Phase 4: Testing: This stage is analogical to evaluation of synthetic data. It includes format correctness testing, values check, parameters check and completion control.

4.2 Examples of semisynthetic benchmark data

4.2.1 Lidar data: Main parameters of the examples of lidar semisynthetic data are presented in table 2.

Parameter	Data 1	Data 2
Name	Krakow Centre	Highway
Kind of data	Block	Strip
Object	City centre	Line object
Deformation	Z	XYZ
Method of deformation	Another method	The same opposite method
Density	12pts/sq.m.	120pts/sq.m.

Table 2. Semisynthetic lidar data.

The lidar semisynthetic data can be extracted from block or strips. Below we present one example of data extracted from block and another from strips. The first is a city centre, the another - corridor data of a highway. The first was deformed by a different process, the second was deformed by the same method as a method to be tested but with the opposite direction. The first was deformed altimetrically, the second in XYZ directions.

1. Krakow Centre data

The first example is a data of Krakow's Market Square area for testing of altimetric correction of the strip data block data (fig. 9). The ISOK (ISOK project, 2011) national project blocks of 12pts/sq.m. were used to extract three overlapping strips of 100m width.

Figure 9. Block of semisynthetic data of Krakow city centre.

The GCPs were arbitrarily selected and marked by red crosses (fig. 10), 8 points per each strip, four in overlap areas. Each GCP can be defined by a specific lidar point that was selected and it is used as a central point of 9 red points added to the strip and featuring the shape of cross.

Figure 10. GCP selected and signed on semisynthetic data.

Next step of preparation of the data was the transformation. The Z coordinates of the GCPs were changed. Each GCP located in overlapping areas got different corrections in different strips. In the next step GCPs were used to deform the strips. The strips were deformed to simulate 3 independent lidar strips using the stereoscopic model deformation method. The a priori Z coordinates were used for correction of the data to the start state. The correction was done by the method to be tested and the result can be compared strictly with height of GCPs and of the original data.

2. Highway data

Corridor data of a highway was chosen as an example data for testing of XYZ correction (fig. 11).

Figure 11. Semisynthetic corridor data of the highway.

In this dataset there are 3 overlapping strips of scanning: the first of 1350m length starting from the north overlaps the second (of 580m length) at a distance of 350m. The second overlaps the last one at a distance of 280m and is 1300m long. The medium density is 120 points per square meter. The point cloud before controlled geometrical deformation was enriched in synthetic GCPs and CHPs. The specific points of the cloud were chosen as a centre of synthetic GCP/CHP. New points in the shape of easy-to-find cross were generated and added to the point cloud (fig. 12).

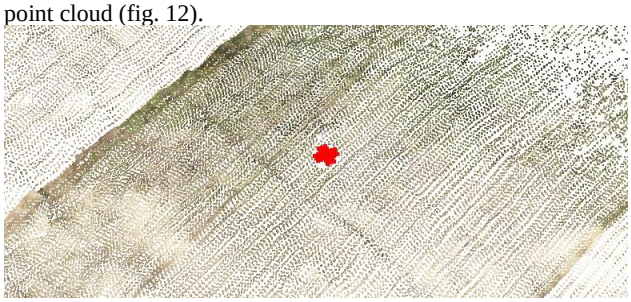

Figure 12. Synthetic sign of GCP.

The 3D coordinates of GCPs were changed and the point cloud was deformed using the lidargrammetric method to be tested later with opposite direction.

4.2.2 Photo data: Photo semisynthetic data that we propose as a raster part of our benchmark dataset is a block of lidargramms generated from real data. The deformation of the block can be done in two ways. The first is to generate lidargramms from deformed lidar data, the second - to change EOPs of lidargramms generated from original, correct data. Both methods of disruption can be used together.

4.2.3 GCPs and CHPs: As it was described above the GCPs and CHPs can be selected on lidar data and signed by crosses of additional points. This method is independent of the specific geometric objects and flexible in location. The correct coordinates of these points are known a priori and they can be changed as much as the data is transformed.

4.2.4 Trajectory: Simultaneous spatial data collection using LiDAR and camera sensors is becoming recently increasingly popular. Such integrated data acquisition became feasible due to commercially available integrated measurement platforms comprised of mapping sensors (scanner, camera), as well as georeferencing systems (IMU, GNSS). While acquiring data, individual measurements (LiDAR points, images) are annotated using georeferencing system time in the form of timestamps (Haala et al., 2022; Pöppl et al., 2023).

Having information about a scanner's trajectory (XYZ coordinates, YPR angles) and offsets between a camera, scanner, GNSS antenna, and IMU systems allows for reducing trajectory errors by considering aligned nodes of images (fig. 13).

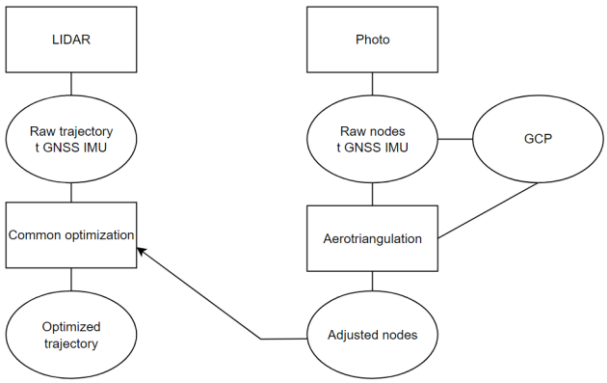

Figure 13. Diagram of integrating GPS, INSS data from Lidar trajectory and Photo nodes. T – time, GNSS – coordinates XYZ IMU – angles Roll/Pitch/Yaw.

Described data structure enables testing various alignment variants both during the aerial triangulation of photos and the subsequent optimization of Lidar trajectories. By utilizing the actual Lidar trajectory, it's possible to artificially generate perturbed trajectories. The perturbation involves adding random or systematic (bias) errors to individual values, both linear and angular (fig. 14).

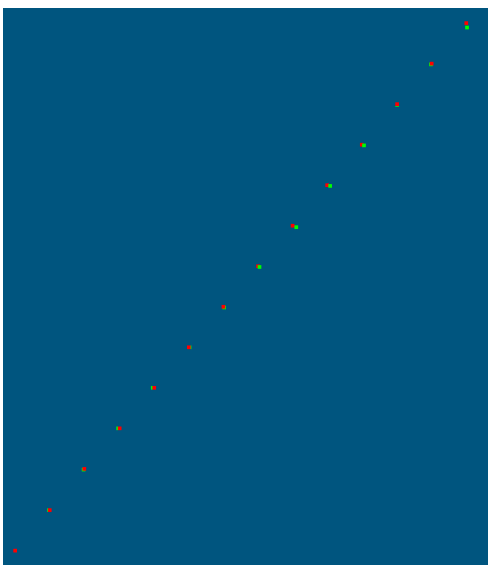

Figure 14. Real trajectory – red, noised trajectory – green.

The set of benchmark data created in this manner comprises various versions of perturbed trajectories, accompanied by corresponding actual values for image nodes. This enables the testing of optimization methods under different perturbation parameters that are hard to simulate during flights.

5. CONCLUSIONS

The collection of a significant volume of real-world data can often be challenging or even impossible. It is also not always feasible to prepare datasets that correspond to various conditions that may occur during measurements. A response to such challenges lies in the generation of synthetic and semi-synthetic data. Presented capabilities and methods of generating synthetic and semi-synthetic data enable the creation of benchmark datasets (comprising point cloud data, photos, GCPs, and trajectory data that have been thoroughly checked and evaluated for testing new methods). In comparison to other datasets of this kind, a novel feature in our dataset is the inclusion of synthetic GCPs and CHPs within LIDAR data.

Currently, a frequently addressed research problem revolves around improving the accuracy of LIDAR data. Research indicates that one possible solution to this problem is the integration of photogrammetric data. Platforms that simultaneously acquire LIDAR and photogrammetric data are becoming increasingly popular. However, obtaining a comprehensive test dataset using such platforms is time-consuming and still does not adequately reflect the various errors that may occur during measurement. The methods we have presented for generating such data synthetically serve as a response to this research problem. They facilitate testing the currently highly developed methods of integrating LIDAR data with photogrammetric data.

REFERENCES

Du, Q., Xie, D., Sun, Y., 2015: An Automatic High Precision Registration Method between Large Area Aerial Images and Aerial Light Detection and Ranging Data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.-ISPRS Arch*., 40, 17–21. doi.org/10.5194/isprsarchives-XL-7-W4-17-2015.

Haala, N., Kölle, M., Cramer, M., Laupheimer, D., Zimmermann, F., 2022. Hybrid georeferencing of images and LiDAR Data for UAV-based point cloud collection at millimetre accuracy. *ISPRS Open J. Photogramm. Remote Sens*, 4, 100014. doi.org/10.1016/j.ophoto.2022.100014.

ISOK. Available online: https://isok.gov.pl/index.html (accessed on 08 September 2023)

Mitishita, E., Costa, F., Centeno, J., 2020: *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci*., V-1-2020, 101–107. doi.org/10.5194/isprs-annals-V-1-2020-101-2020.

Parmehr, E.G., Fraser, C.S., Zhang, C., Leach, J., 2013: Automatic Registration of Optical Imagery with 3d Lidar Data Using Local Combined Mutual Information. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*., 2, 229–234. 10.5194/isprsannals-II-5-W2-229-2013

Pöppl, F., Neuner, H., Mandlburger, G., Pfeifer, N., 2023: Integrated trajectory estimation for 3D kinematic mapping with GNSS, INS and imaging sensors: A framework and review. *ISPRS J. Photogramm. Remote Sens.*, 196, 287-305. doi.org/10.1016/j.isprsjprs.2022.12.022.

Rodríguez-Cielos, R., Galán-García, J.L., Padilla-Domínguez, Y., Rodríguez-Cielos, P., Bello-Patricio, A.B., López-Medina, J.A., 2017: LiDARgrammetry: A New Method for Generating Synthetic Stereoscopic Products from Digital Elevation Models. *Appl. Sci*., 7, 906. doi.org/10.3390/app7090906.

Rzonca, A., Twardowski, M., 2022: The lidargrammetric model deformation method for altimetric UAV-ALS data enhancement. *Remote Sensing*, 14, 6391. doi.org/10.3390/rs14246391.

Schofield, S., Bainbridge-Smith, A., Green, R., 2022: An improved semi-synthetic approach for creating visual-inertial odometry datasets. *Graphics and Visual Computing*, 200061. doi.org/10.1016/j.gvc.2022.200061.

Valbuena, R., Mauro, F., Arjonilla, F.J., Manzanera, J.A., 2011: Comparing Airborne Laser Scanning-Imagery Fusion Methods Based on Geometric Accuracy in Forested Areas. *Remote Sens. Environ*., 115, 1942–1954. doi.org/10.1016/j.rse.2011.03.017.

Wang, F., Zhuang, Y., Gu, H., Hu, H., 2019: Automatic Generation of Synthetic LiDAR Point Clouds for 3-D Data Analysis. *IEEE Transactions on Instrumentation and Measurement*, 68(7), 2671-2673. doi.org/10.1109/TIM.2019.2906416.