Semantic generalization of the multi-source digital elevation model (DEM)

Tomaž Podobnikar

Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, Ljubljana, Slovenia;
tomaz.podobnikar@fgg.uni-lj.si

Abstract: The aim of the study is to demonstrate a utilization of generalization in a complex task of multi-source data conflation. The data conflation means using existing data sources of often dissimilar quality to expose their best properties in the final product. The digital elevation model (DEM) from multiple existing data sources is topical when a large quantity of spatial data is available, what makes better DEM quality feasible. The main focus is assigned to the following selected aspects of generalization: (1) extensions of the standard generalization, (2) characterization and homogenization of the data sources, (3) the conception of the target DEM, and (3) quality control/cleaning. We recognized the importance of the generalization for general, in a contrast to specific purposes, such as for cartography and spatial analysis. The generalization is also an important part of homogenization of data sources for interoperability, filtering and interpolation into continuous surfaces, and for multi-scale DEM processing. The most practical part of the study proposes employment of the proposed generalization solutions in the pre-processing and processing phases of the DEM conflation. We tested solutions for the generalization that includes well-known geometrical, as well as semantic – mainly geomorphological properties (semantic signatures/DEM parameters, semantic indicators). Especially the proposed generalization, which involves a semantic knowledge, needs a support of the expert interpretation, reasoning and interaction. The studied aspects are highly sustainable in terms of the extended lifecycle of the datasets and quality enhancement.

Keywords: Digital Elevation Model, DEM, Generalization, Semantic Multi-Source Data Conflation, Data Integration

1. Introduction

The main objective is to study various aspects of generalization of a digital elevation model (DEM), processed from multi-source datasets. The DEM, often called as digital terrain model (DTM), was one of the first forms of digital geographical information that became available (Fisher and Tate 2006), which is now a fundamental dataset to mapping our planet. It is considered to be amongst the most important datasets in the geospatial domain for the greater part of spatially oriented applications and research.

Our challenge is that the volume of DEM datasets is rapidly increasing due to a long-term explosion of information with exponential growth (Price 1963). Nearly invariably, the information for DEM processing comes from a large single source dataset. However, user demand for higher quality outputs and increasing levels of detail is starting to reveal the limitations of this single source approach. The DEM from multiple existing data sources is topical when a large quantity of digital spatial data is available, what makes better DEM quality feasible.

The abstraction with a conceptualization of the DEM, and consequently the generalization is an important process in order to exceed the common inhomogeneity of the DEM from multiple sources. This process can consider any user requirement. An additional challenge is applying various aspects of semantics that are rarely taken into account for the geospatial datasets generation. The DEM is namely a complex model that includes inherently recorded information, which can be used for such generalization.

1.1 Multi-source DEM

Heterogeneous data sources have a potential to be combined in order to enhance and solve tasks. We recognized three typical options. The first option is spatial data integration process, where the larger set is retained, i.e. databases are merged (Butenuth et al. 2007). An example is the implementation of spatial ETL (extract, transform, load).

The next option is spatial data fusion that combines heterogeneous data sources (and knowledge) into a consistent, accurate and useful representation. With the fusion, the quality of data is improved, but the final product is often considerably different with regard to the source data. A typical example is data from different sensors used to produce better raster images, e.g. with pansharpening (Ehlersa et al. 2010), or a comprehensive gemorphometric approach based on a synergy of satellite images and DEM, in order to enhance analytical shading.

The third option is a spatial data conflation. It is similar to the fusion, but in this case is important that the quality of the resulted dataset is better than any input individual DEM (source) for the target area. The data conflation means using existing data sources of different quality (or even dirty data) to expose their best properties in the final product. The example is a conflation a number of different DEMs into a single DEM of considerably better quality (Tang et al. 2014).

The framework of this paper is based on the third option – spatial data conflation – more particularly on multi-source and optionally multi-temporal spatial data conflation, called progressive Weighting Sum with Geomorphologic Enhancement (WSGE) method (Fig. 1) and Synchronic Interpolation (SI) method. Principles of the both semi-automated methods were initially developed and described in Podobnikar (2005). Some of these solutions were already implemented in the national DEM of Slovenia and in other applications. The essential issue with multi-source DEM (big geospatial data) is the amount of data, different types of data sources, different quality standards, inconsistency, and nevertheless various rates of data generalization.

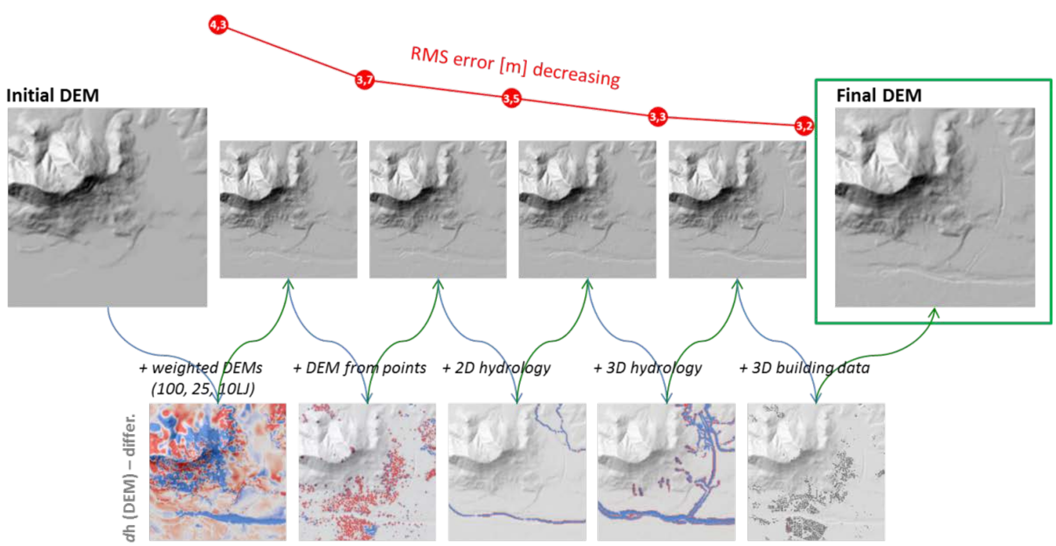


Fig. 1. A DEM conflation presentation with a WSGE method, from initial to the final DEM.

1.2 DEM generalization

The generalization is essentially connected with DEM conceptualization (abstraction). The main concept of generalization is based on the implementation of a less-specific criterion to the target model. In our case, the aim is to produce a smaller scale DEM with the coarser resolution from a larger one, with changing its granularity. This process is called downscaling in some contexts. There are different types and views of the DEM generalization. A (spatial) geometrical generalization (reduction, displacement, merging, etc.) simplifies an object's geometry. A semantic generalization classifies objects according to their properties and a related topological (and contextual, geomorphological) generalization changes the relations among objects correlated with geometry in the cartographical discipline. A temporal generalization considers multi-temporal aspects of the terrain development through the time, which is in our case connected with the homogenization of the DEMs, similar to the multi-source DEM processing.

2. Generalization in the context of multi-source DEM

The multi-source DEM production is divided into following subsequent phases: preparation, pre-processing of datasets and processing (conflation) into the final DEM (Podobnikar 2005), whereas the generalization aspects that appear in the last two phases are studied. We propose the following aspects (Fig. 3):

1. Extensions of the standard generalization
2. Characterization and homogenization of the data sources
3. Conception (formulation) of the target DEM
4. Quality control and cleaning

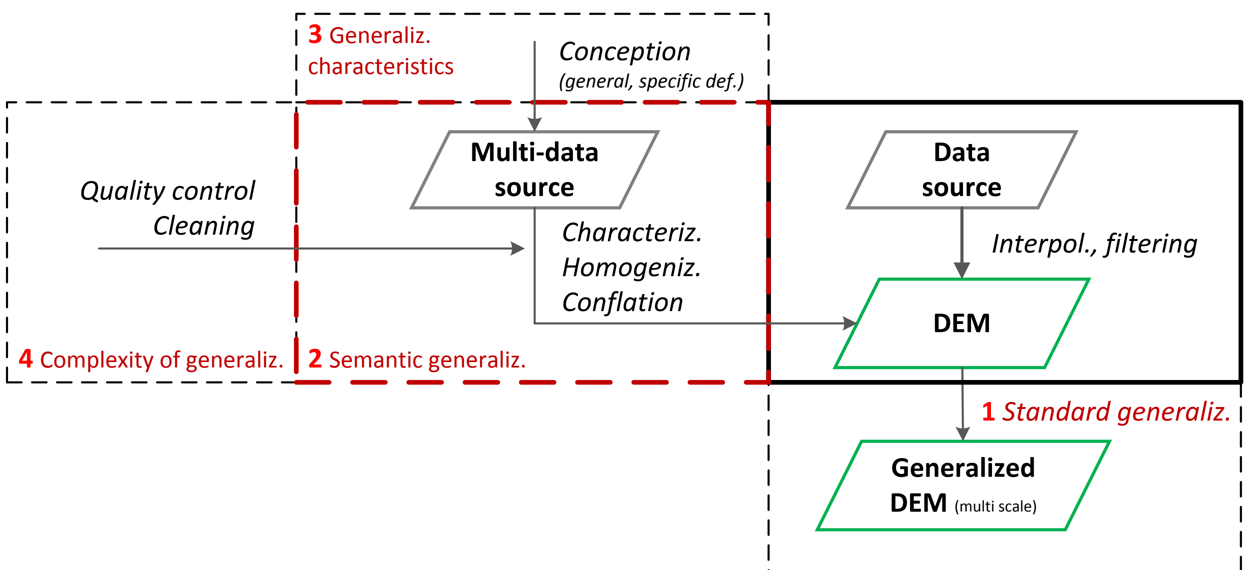


Fig. 3. Generalization types in the context of the multi-source DEM processing, where a typical DEM generation procedure is presented in the non-dashed frame, while the core multi-source pre-processing and processing are in the red frame.

2.1 Extensions of the standard generalization

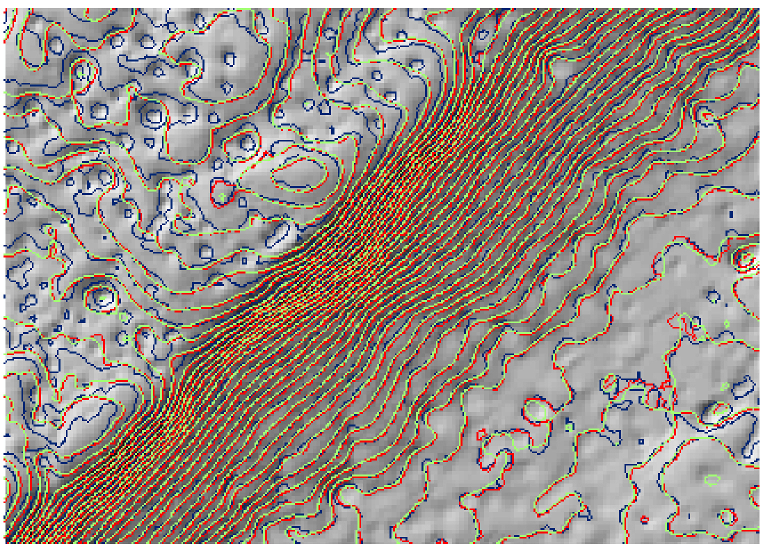
The so-called "standard generalization" means a generalization of the already generated DEM. The following extended aspect of the DEM generalization is proposed and developed:

- Generalization for a general use (universal, multi-purpose)
- Generalization for selected specific purposes:

(A) Cartographic presentation
(B) Spatial analysis in geographical information systems (GIS)

For the general use of a generalized DEM, the greater number of conditions for specific purposes is reasonably considered. In the case of a conflated near universal DEM, it is needed optimally geometrical and geomorphological properties of the datasets in any target scale and time.

Two essential specific purposes according to user requirements, including social aspects have been discussed in a greater detail. The cartographic presentation (A) typically applies contour lines (Fig. 2), analytical shading or hypsometry. Particular in cartography are cartographic scale conceptions, instead of the resolution of the DEM, and a specific cartographic generalization. Moreover, the mapping rules are used to generate a valid map. For the cartographic DEM production, the geomorphological inaccuracy is, therefore, a less tolerated than the positional one. Different kinds spatial analyses in GIS (B) using a DEM can cause unexpected results. Some analysis can generate significantly different results on just slightly differently generalized DEM. Depending on the analysis, positional or/and geomorphological appropriate quality DEM is required. For example, altitudinal zonation analysis is, in general, less sensitive to the DEM quality than computing a slope. On the other hand, a robust spatial analysis or variables can successfully tolerate certain types of inaccuracies and uncertainties in the DEM. The error propagation methods, such as Monte Carlo simulation can help understanding the sensitivity of a particular generalization approach.



Contour lines from
National DEMs:

Green: DEM 12.5 m
Red: DEM 5 m
Blue: DEM 1 m

Area of size 1.5 x 0.8 km

Fig. 2. Which DEM is better for this scale to generate the standard contour lines? (karstic test area in Slovenia)

An option is that the final product would be available as a multi-scale DEM. For the general use, our strategy is to generate a DEM with the double resolution and then applying a combination of appropriate interpolation and resampling methods, with stepwise downscaling. For example, when the target set of multi-scale DEMs were in the resolution of 12.5, 25 and 100 m, the initial DEM should have a resolution of 6.25 m. This high-resolution DEM is then stepwise generalized into a set of target DEMs of coarser resolution.

The success of a particular generalization method of DEM in combination with interpolation of vector data sources for DEM generation, such as contour lines, depends on how these sources are sampled, on their geomorphological appearance and accuracy, and especially on the chosen interpolation method. Our strategy for a multi-source DEM combines different methods for interpolation/filtering and weighting based on actual geomorphology. This approach can be also applied to the individual DEMs (e.g. Podobnikar and Vrečko 2012).

2.2 Characterization and homogenization

2.2.1 Characterization and parametrization of the data sources (semantic signatures, weights)

From semantical view of the DEM generalization, the methods for detection and analysis of shapes with characteristic features and landform, e.g. mountains and peaks (Podobnikar 2012), karst depressions (Obu and Podobnikar 2013), talus slopes (Podobnikar and Székely 2015) and the entire terrain skeleton in various scales/resolution (Zakšek and Podobnikar 2005), the automated detection of genetic types (stone structure), soil coverage or types, degree of karstification; vegetation cover, etc. were developed (Podobnikar 2010). In addition to the environmental, a number of anthropogenic characteristics were modelled, such as transport network, stone quarries, etc. For example, it is well known the highest inclination of the road network is 12% or 30%, and of the railways 12‰ to 25‰), depending on their category.

Semantic signatures are considered as kind of variables in individual DEM characterization used for the conflated DEM processing, which can help to uncover implicit information from the data, i.e. to assess and delineate the target features/parameters of the surface. The analogy to the proposed approach is spectral signatures used for supervised classification in remote sensing (Janowicz 2012). The parameters in Table 2 were used for segmentation to regions, weighted for individual DEMs and implemented into a decision-making problem-solving procedure (Forman and Selly 2001), in order to produce surfaces used for conflation. The order of data sources for conflation is set at the end.

Table 2. Target DEM parameters.

DEM parameters **Level 1**	**Level 2**	***Example: DEM 10*** ***Action***	***Weighting [0-100])***
Geometric indicators	terrain skeleton (vector) in different scales, contour lines	*peaks, sinks, structural lines*	*70*
Semantic indicators (natural)	geomorphological properties	*slope, roughens*	*50-100*
	complex features description (shapes)	*a DB of peaks, karst depression*	
	genetic types (including degree of karstification)	*4 categories according to surface roughness*	*<100*
	soil cover	*2 categories*	*<100*
	vegetation cover	*4 categories (photogram. DEM)*	*<100*
	temporal changes – absolute (landslide)	*landslide from 2016 identified*	*100*
	temporal changes – relative (earthquake, mining, long term landform development)	*recent land subsidence detected by InSAR – modelling*	*model implement.*
Semantic indicators (anthropogenic)	transport network, stone quarries, rubbish dumps	*a DB of the stone quarries*	*100*
	road and railway network	*a DB of the roads*	*<100 (model implement.)*
Propose of use	specific, general	*general*	*for standard implem.*
	which features are applicable for a target time frame	*exclude stone quarries*	*model implement.*
Resolution	scale, grain, grid size	*for resolution 5 m*	*for concept implem.*
Quality	statistical	*(RMSE, max)*	*for weighting implem.*
	empirical, visual	*analytical shading*	*for elimin. errors*

There are different solutions for parametrization of geodetic data points of different density, where the point can be considered as an individual feature, while the set of points as a surface.

Additional semantic information is acquired from other geospatially related data using different methods, e.g. by metadata and acquired through spatial data mining. Other challenges are cross-domain knowledge integration and getting reliable outputs (concerning scalability and flexibility) for applications that can satisfy various key customer' needs.

The described procedure is useful for mixed, geometrically and geomorphologically conflated DEM. Namely; the DEM conflated with weighing is typically geomorphologically inhomogeneous. Therefore, an important step is a selection of geomorphologically best data source that is applied at the end of DEM processing with a trend surface method. The result is then a geometric-geomophologic optimally conflated DEM using semantic principles, in particular resolution (through semantic generalization).

2.2.2 Homogenization of data sources for interoperability

There is no unique solution to the DEM generalization, and there is no unique problem to solve, e.g. generalization of contour lines vs. scattered points; or homogeneous data sources vs. disparate data sources. An additional human reasoning is important to get a high-quality and reliable result. The generalization principles can help in the geometric and semantic homogenization (matching) of the various DEMs for multi-source conflation. These DEMs as data sources are typical of different resolution, acquired and modelled with different methods (e.g. photogrammetrically or from Lidar), based on different standards (e.g. generated as digital surface models (DSM) with objects or as well defined DEMs).

There are a number of approaches in order to analyse and homogenise the data sources. We expose the main approaches applied for:

- Upscaling (homogenization to a highest optimal resolution; using parameters from Section 2.2.1): resampling, predictive and interpolation methods on grids; subpixel classification to get characteristic structural points; enrichment with spatial data sampling to enhance the terrain skeleton (e.g. according to Makarovič 1973, Podobnikar 2005).
- The other are approaches for downscaling: reduction the (big) data in order to fulfil geometric and semantic criteria for generalization (in the case of oversampling), e.g. to the terrain skeleton, resampling, smoothing, filtering.

2.3 Conception of the target DEM

It is an ontological problem to present formally a concept of a DEM as a finite (generalized) model of infinite and countable reality. The principles of conception and semantics are crucial for semantic generalization of the multi-source DEM. The standard conception of the DEM had been enriched with an analytical description of anthropogenic and environmental characteristics of terrain for the regional and global areas and corresponding grid size (Table 1).

Table 1. Conceptualization with proposed semantic indicators.

Indicator	***Example***
conception, definition, standards for DEM	*excluding vegetation, snow cover, buildings (houses and bridges); the surroundings of topographic heights not need extreme attributes*
purpose of use (multi-purpose = general use, and specific-purposes)	*for cartography – contour lines*
time frame (paleo, contemporary, etc.)	*1995–2000*
automated generalization, (level of detail (LoD), scale/grain/grid size, etc.)	*scale 1:250,000, grid size 50 m*
geomorphologic/hydrologic characteristics (based on slope, roughness, runoff, etc.)	*geomorphologically detailed, hydrologically with medium detail*
quality information	*allowed low RMSE of 50 m in mountain area, 5 m in plains*

2.4 Quality control and cleaning

The quality assurance and cleaning are very complex topics in spatial information applications where it is challenging to set independent, objective and robust methods. The complexity of the multi-source DEMs quality can be substantial when the observed phenomenon is far from trivial, e.g. in specific geological settings, or when the corresponding datasets present the DEM phenomena in multi-scale or multi-temporal modes. The complexity of the differently generalized DEM quality can be, for example, measured with information entropy (Hu et al. 2015).

Our conflation approach to generate a multi-source DEM considers geometrical (position), temporal and topographic (shape, semantic) dimensions of quality. The last one is most demanding to measure. However, which accuracy is more important? Is it a geometrical accuracy that applies heights of the mountain peaks to a DEM (to the individual grid cells), or is it the semantic accuracy that applies geomorphologically correct shapes (to a set of grid cells)? The last is mostly neglected in the practical solutions. For both, general and particular purposes of use, our solution of the data conflation considers a kind of weighting between geometrical and semantic accuracies to assure optimal DEM quality. Therefore, a value of the individual grid cell in any scale is never calculated as a simple average of the terrain elevations within them. The multi-source DEM generalization involves error characterization and cleaning procedure for gross, systematic and locally systematic error with a proposed empirical Spatial Error Model (SEM) (Podobnikar 2016).

4. Conclusions

This paper demonstrates a comprehensive utilization of generalization for a complex task of multi-source data conflation. It also raises a number of questions that are specific for multi-source DEM conflation. Acquiring information about semantic properties is crucial, but far from trivial in relation to generalization, as well as in relation to (dirty) big data sources, conceptualization, interpolation, spatial quality, etc.

We exposed some aspects towards the optimal solution in a pre-processing and processing phases of multi-source DEM conflation aiming generalization. We identified, proposed, and partly tested solutions for the generalization that comprises geometrical and semantic (geomorphological) properties. We also recognized the importance to develop solutions for the generalization in the case of general and selected specific purposes for cartography and spatial analysis. The generalization was also an important part of homogenization of data sources for interoperability, filtering and interpolation into continuous surfaces and for multi-scale DEM processing.

Future idea based on the development of multi-data sources DEM ontology in relation to the universal ontology of geographic space (e.g. Podobnikar and Čeh 2012). This will cover some already discussed topics, such as semantic enrichment and homogenization, interoperability, data conflation process, as well as a development of the semantic reference system (Kuhn 2003), measures for semantic similarities, etc.

The main limitation of the proposed generalization issues is that a fully automated processing chain is difficult to reach. Especially the generalization process that involves the semantic knowledge still needs a support of the expert interpretation, reasoning and interaction. However, the proposed aspects allow better control of geospatial solutions and are highly sustainable in terms of the extended lifecycle of the datasets and quality enhancement.

Acknowledgements

Part of the study was supported by the Water Science and Technology and Geotechnics Program (P2-0180) of the Slovenian Research Agency. The author would also like to thank the anonymous reviewers for their comments.

References

Butenuth, M., Gösseln, G.v., Tiedge, M., Heipke, C., Lipeck, U. Sester. M. (2007). Integration of heterogeneous geospatial data in a federated database. *ISPRS J Photogramm Remote Sens*, 62(5), 328–346.

de S. Price., D.J. (1963). Little science, big science. New York: Columbia University Press.

Ehlersa, M., Klonusa, S., Åstrandb, P.J., Rossoa, P. (2010). Multi-sensor image fusion for pansharpening in remote sensing. *Int J Image Data Fusion*, 1(1): 25–45.

Fisher, P.F., Tate, N.J. (2006). Causes and consequences of error in digital elevation models. *Prog Phys Geog*, 30(4), 467–489.

Forman, E.H., Selly, M.A. (2001). Decision by Objectives: How to Convince Others that You are Right, *World Scientific.*

Hu, L., He, Z., Liu, J., Zheng, C. (2015). Method for Measuring the Information Content of Terrain from Digital Elevation Models. *Entropy*, 17, 7021–7051.

Janowicz, K. (2012). Observation-Driven Geo-Ontology Engineering, *T GIS*, 16(3), 351–374.

Kuhn, W. (2003). Semantic reference systems. *Int J Geogr Inf Sc*, 17(5), 405–409).

Makarovič, B., (1973). Progressive Sampling for Digital Terrain Models. *ITC Journal*, 3, 397–416.

Obu, J., Podobnikar, T. (2013). Algorithm for karst depression recognition using digital terrain models. *Geod. vestn.*, 57(2), 260–270.

Podobnikar, T. (2005). Production of integrated digital terrain model from multiple datasets of different quality. *Int. J. Geogr. Inf. Sci.*, 19(1), 69–89.

Podobnikar, T. (2010). DEM Production/Updating Based on Environmental Variables Modeling and Conflation of Data Sources. *J Civil Engineering and Architecture*, 4(11), 33–44.

Podobnikar, T. (2012). Detecting Mountain Peaks and Delineating Their Shapes Using Digital Elevation Models, Remote Sensing and Geographic Information Systems Using Autometric Methodological Procedures. *Remote Sens.*, 4(3), 784–809.
Podobnikar, T. (2016). Combination of error characterization and spatial error model to improve quality of digital elevation models. Accuracy 2016, 241–248.
Podobnikar, T., Čeh, M. (Eds.) (2012). Universal Ontology of Geographic Space: Semantic Enrichment for Spatial Data, *IGI Global*, 276 p.
Podobnikar, T., Székely, B. (2015). Towards the automated geomorphometric extraction of talus slopes in Martian landscapes. *Planet. Space Sci.*, 105, 148–158.
Podobnikar, T., Vrečko, A. (2012). Digital Elevation Model from the Best Results of Different Filtering of a Lidar Point Cloud. *Trans. GIS*, 16(5), 603–617.
Tang, Y., Zhang, J., Li, H., Ding, H., Liu, Q., Jing, L. (2014). A multiple-point geostatistical method for digital elevation models conflation. *IGARSS*, *IEEE*, 4299–4302.
Zakšek, K., Podobnikar, T. (2005). An effective DEM generalization with basic GIS operations. 8th ICA Workshop on Generalisation and Multiple Representation. Coruńa, 10 p.