The 9th International Conference "ENVIRONMENTAL ENGINEERING"
22–23 May 2014, Vilnius, Lithuania
SELECTED PAPERS
eISSN 2029-7092 / eISBN 978-609-457-640-9
Available online at *http://enviro.vgtu.lt*

Section: Technologies of Geodesy and Cadastre

User perspective on geospatial data quality. Case study of the Polish Topographic Database

Elzbieta Bielecka[a], Małgorzata Leszczyńska[b], Peter J Halls[c]

[a]*Military University of Technology, Faculty of Civil Engineering and Geodesy, 2 Gen. S. Kaliskiego St., 02-908 Warsaw, Poland*
[b] *University of Warmia and Mazury in Olsztyn, Faculty Geodesy and Land Management, Oczapowskiego 2, 10-957 Olsztyn, Poland*
[c] *University of York, Derwent College, Reconstruction and Development Unit (PRDU), Heslington, York YO10 5DD*

Abstract

The aim of the paper is to analyze the quality of geospatial data from the user perspective, which is application oriented and differs from that of the data provider due to the performance of specific tasks. Generally the overall description of data quality, reported in metadata, comprises characteristics of the data set parameters such as completeness, accuracy, and consistency of data. However, this information is very important it shows data conformance with its specification but for most of the users is not sufficient. The authors present innovative method for assessment data quality on the level of object classes focusing on spatial location of missing objects and their attributes as well as the information whether some attributes assume null reason values. The results are obtained automatically, on the basis of elaborated algorithm and reported in the form of choropleth map which should be displayed as one of the layer in a geoportal. The method was tested on land cover dataset, stored in the Database of Topographic Objects for Olsztyn.

Keywords: geospatial data; data quality; metadata; choropleth map.

1. Background

Since the mid-1990s concerns about geospatial data quality have increased as a result of the digitisation of resources previously collected as maps and the development of spatial data infrastructures, which make spatial data available via network services. Both data producers and users have become increasingly concerned about their inability to measure and communicate the quality of geospatial data or other geoinformation products. It has also been observed that there are more non-expert users who use data for multiple purposes [1] who should be informed of data incompleteness or inconsistency. Generally, the quality of dataset is described in metadata, in a standardized way, according to ISO19 115, Dublin Core standards or INSPIRE metadata regulation and implementation rules. Data quality is assessed against ISO 19113 quality principles, ISO 19 114 – quality evaluation procedures and ISO 19138 – quality measures. Metadata provides information on several aspects of the datasets, such as identification, spatial reference system, geographical location, temporal reference, data quality and validity, constraint related to access and use, as well as organizations responsible for the establishment, management, maintenance and distribution of spatial data sets and services. Metadata, however, are prepared by data producers who describe the set with regard to its conformance with data specifications. Since user requirements are connected with the performance of a specific task, the data description provided by the producer does not always meet user needs [2–3].

Data quality issues have been extensively explored since the mid-1990s. During the last 25 years of research, two trends have been observed in the literature: the first – describing the intrinsic characteristics of the dataset, resulting from data production methods (conformity with data specification, data format, spatial resolution, i.e.), and the second – where quality is defined as the level of fitness between data characteristics and user needs, is often identified as external quality [4]. The fitness for use data quality concept is relative to users needs [2] and is neither independent nor absolute. Many researchers [3], [5–7] argue that quality assessment described as fitness for use requires data characteristics that are not yet included in geospatial metadata standards. Recently, the focus of geospatial data quality research has shifted from developing more complex methods of measuring quality to finding out what kind of quality information users actually need to avoid data misuse [8–10]. Although there have been a considerable number of studies related to evaluating and describing geospatial

Corresponding author: Małgorzata Leszczyńska. E-mail address: malgorzata.leszczynska@uwm.edu.pl
http://dx.doi.org/10.3846/enviro.2014.193

data quality [1–7, 9–10], as well the influence of data quality on results of spatial analysis [11–13], there is no research on data incompleteness in the sense of missing or voidable objects or their attributes. From a user's perspective, besides the general description of completeness, accuracy and consistency of data, knowledge of missing objects and their attributes is also essential as well as the information on whether some attributes assume null reason values, which de facto means a lack of information. This research seeks to fill this gap, illustrated in the form of a choropleth map showing the differences in quality within the dataset and showing by a visual variable – color lightness – places where the quality of data set differs.

2. Methods and data

2.1. Method of data quality evaluation

The method of user data quality assessment, also called an "external assessment", is dynamic in character, because it is created automatically based on a user's query. It can also be made either offline or online (remotely), where data are made available in a client-server architecture. If an evaluation is made offline, GIS software is necessary, enabling the selection of objects by attribute query and to attach the results to the thematic layer containing the mapping unit. When an evaluation is made online, it is necessary to develop a special application embedded in the geoportal, which enables a user to select the evaluation criteria and parameters. Such criteria may concern both the thematic scope (selection of the layer being evaluated and the attributes) and the spatial scope, by providing the coordinates of the bounding box or by indicating some polygon. An evaluation may be made in regard to such items as completeness, geometric accuracy, topological correctness and other characteristics. The results can be displayed on a screen as a thematic map and recorded in metadata. However, recording the results of a quality evaluation in metadata which are created in real time based on a user query requires the development of an appropriate profile for metadata and the application which updates the metadata automatically

This paper evaluates the quality of spatial data in terms of completeness of optional data. Individual stages of the spatial data quality involved:

- selection of the quality measures;
- establishing a reference unit – the minimal mapping unit, to which dynamic values of quality measures will be aggregated;
- the development of a quality assessment algorithm;
- the choice of a method of presentation of the quality assessment results.

After preliminary analyses, it was decided that:

- the quality measure is the number of objects which fail to meet the criterion of completeness of values of optional attributes;
- a grid with a cell size of 500 m is the minimal mapping unit;
- the quality assessment results are published as a choropleth map, on which the percentage of objects with missing attributes values are presented with a visual variable – color lightness.

The choice of the size of the reference unit was based on: the size of area under analysis, the number of objects being assessed and limitations of the use of computer technology for cartographic presentations. The values obtained by the mathematical calculations of the cell size were reduced to the range in which a grid is well-visible on 14-inch screens. The proposed size is an average size of screens at stationary units and displays of mobile devices (laptops, tablets and smartphones).

An algorithm for a data quality assessment made offline was developed in the SQL3 language for database handling. Due to the non-procedural features of the language, GeoMedia 6.1 software was used to execute commands. The software also enabled the aggregation of selected objects to reference units (grid cells) and visualisation of the results on a map. The SQL3 language also makes it possible to show a result in a grid cell and to adapt the method of storing attributes to the need associated with preparing an outcome map. The values of the quality measure are converted to percentage values and, subsequently, to sequences of characters, owing to which the percentage values can be displayed on a map. This solution helps to present the geospatial data quality in a generalised manner as a label assigned to a grid cell. Owing to such a record, a user can make quick and clear assessments, without risking a wrong interpretation of the cell value.

2.2. Data

The Database of Topographic Objects in Poland is maintained and distributed by National Mapping Agency -Head Office of Geodesy and Cartography. This is a spatially continuous, vector database with the thematic scope and a level of detail corresponding to contemporary, civilian topographic maps on the scale of 1:10 000. Information on land cover, defined as a description of the surface of the earth by its (bio-) physical characteristics, is one of the 11 thematic layers. Land cover nomenclature comprises 10 classes, such as: built-up areas, road, rail and airports areas, non-built-up areas, arable lands and

pastures, permanent crops, woodlands and shrubs, shrubby area, water bodies, bare lands, non-developed areas [14]. Each land cover class (feature type) is geometrically represented as polygon and characterised by a set of attributes.

The investigation of data quality has been performed for woodlands and shrubs object classes (elaborated according to xsd BDOT schema) which is a part of Land Cover thematic layer. The data covers Olsztyn city and its surroundings area, located in northern Poland (Fig. 1)

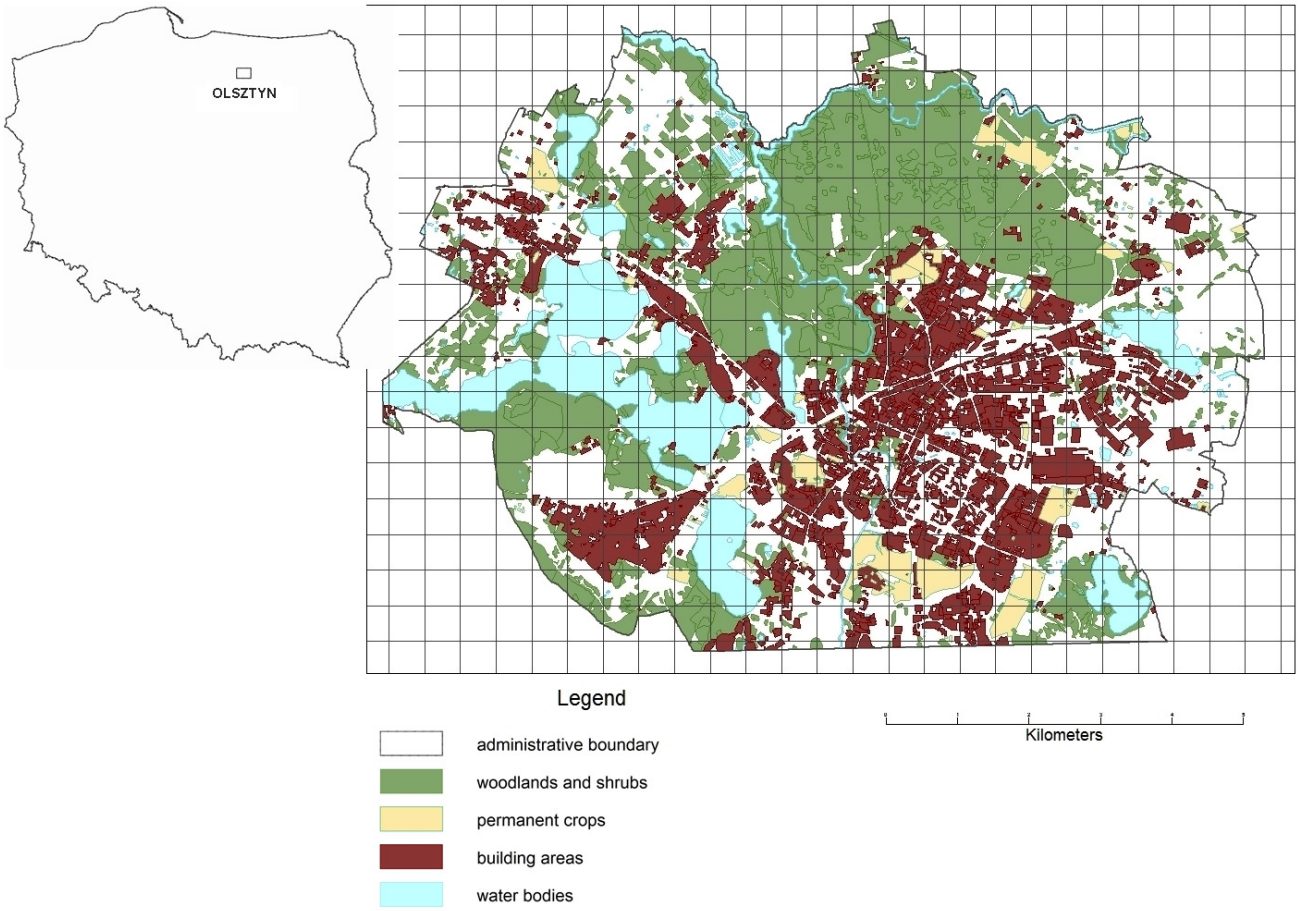

Fig. 1. Analysed data location of pilot area and visualization of woodland and shrubs thematic layer

These data were updated in 2012, mainly on the basis of orthophotomaps, field reconnaissance and the data from large-scale databases (basic map, land register maps). The total mean error of an object location does not exceed 5 m. The overall characteristics of the analysed data are given in Table 1.

Table 1. Descriptive characteristics of *woodland and shrubs* object class

Item name	Value
Geographical extent:	20°22′07′′ W; 20°34′00′′ W 53°49′57′′ N; 53°44′46′′ N
Area covered [in hectares]	2358.1
Updateness	2012
The RMS of object location[in meters]	5
Total number of objects	371

3. Results and Discussion

The analysis results are presented in Table 2 and in Figure 2 and 3. The number of objects with incomplete attribute value, hereinafter referred to as low quality objects (LQO), is 85, which constitute 22.9% of total *woodland and shrubs* objects. In 46 gridd cells the number of LQO is one, it is 12.4% of total number of LQO. More than 5 LQO is just within 4

grid cells, maximum number of LQO is six (only one grid cell). Analysis of Table 1 and Table 2 show that in majority of grid cells (77.1%) all objects have assigned all optional attributes values.

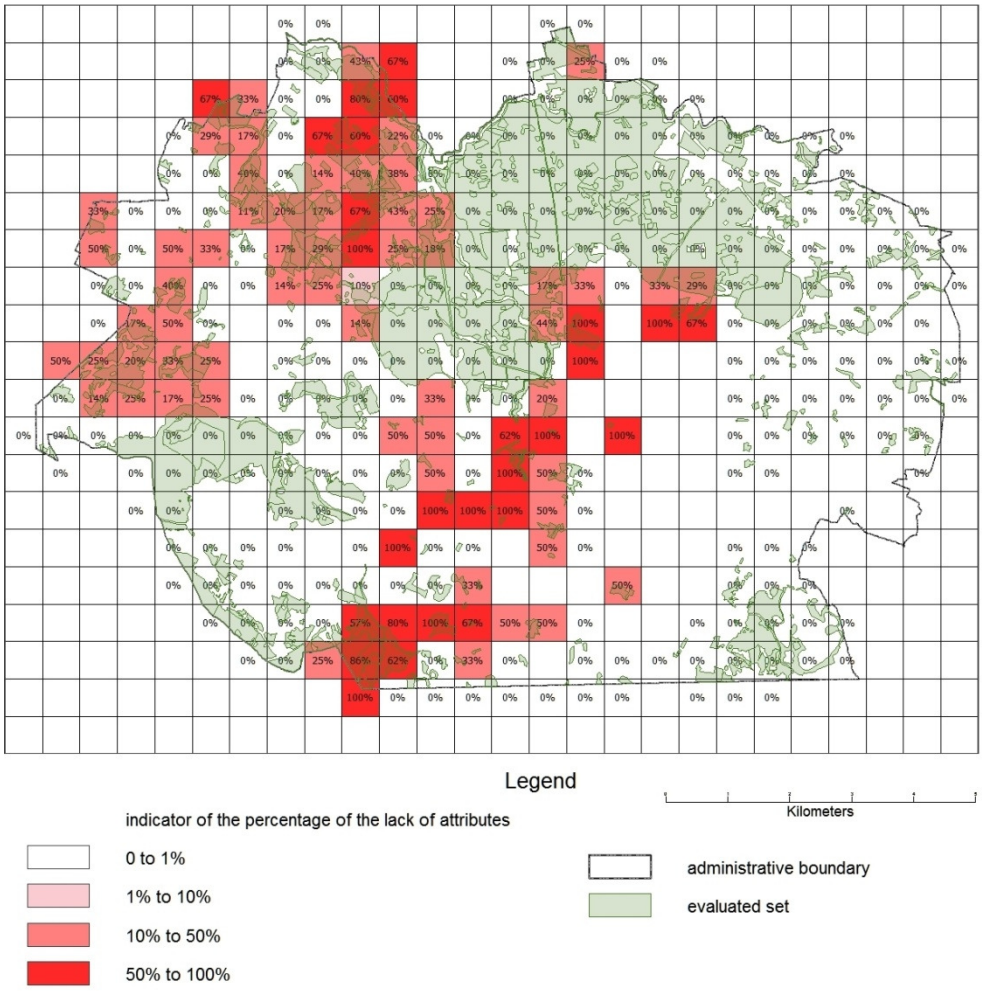

Fig. 2. Percentage of woodland and shrubs polygons with incomplete attribute values

Table 2. An example of a table

Number of LQO	No of grid cells with LQO	% of grid cells with LQO
0	286	77.1
1	46	12.4
2	21	5.6
3	6	1.6
4	8	2.2
5	3	0.8
6	1	0.3
Total	371	100

Figure 3 shows the different approach – the percent of objects with complete attributes, so called high quality objects (HQO). The grid cells are shaded in proportion to decreasing number of HQO, i.e. the darkest colour shows the grid cells with all complete data, while the lightest – those in which optional attributes have the NULL value. A preliminary analysis of perception of both maps shows that users prefer the one which shows the percentage of high quality objects.

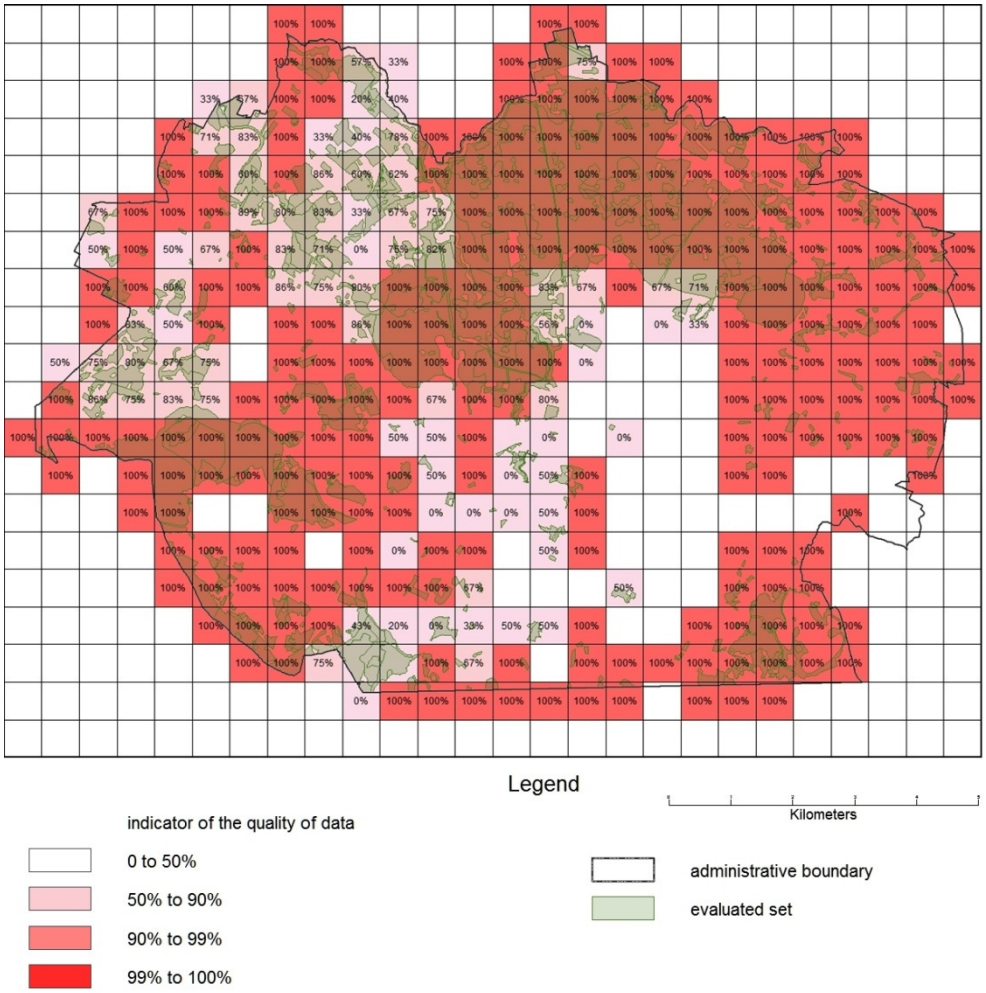

Fig. 3. Percentage of woodland and shrubs polygons with complete all attribute values

4. Conclusions

Although standardised data quality indicators, developed by data producers, are very important, users contend that this kind of description, however valuable, is not entirely sufficient to meet their needs. Any application will have requirements for some level of data quality and it is necessary to know whether the available data will support the application. The method described, which is automatic and easy to develop, can be used to assess the quality of geospatial data and compare the quality of several data sets.

This method of data quality assessment is addressed to users of geospatial data and, due to its relatively simple computational algorithm, based on SQL query, it can be used by even inexperienced users. The results of a quality assessment are presented as a thematic map, which could be published via WMS or WFS network service and used together with the data being assessed. Results could also be saved as a separate set of geospatial data, which enables further data analysis.

The measure values are dynamic, because they change as the data set is updated, but they also change with a user's changing expectations.

The authors regard it as reasonable that such measures should be included in the set of metadata.

The evaluated *woodland and shrubs* object class of Land Cover data, stored in Database of Topographic Objects is of high quality. Although the number of grid cells with missing attribute values is nearly 23% they are assigned to one, big woodland complex.

As standardized form of visualization quality evaluation results with a predetermined number of classes and their ranges, enables the comparison of the quality of several sets or analysis of whether and how the quality of the geospatial data set changes.

Acknowledgements

This paper was prepared as an outcome of statutory research no. PBS/854/2013, conducted at the Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geodesy, and as an outcome of statutory research no. 528-0302-0828 Faculty of Geodesy and Land Management, Institute of Geodesy.

References

[1] Gervais, M. 2006.On the importance of external data quality in civil law, in R. Devillers, R. Jeansoulin (Eds.). *Fundamentals of spatial data quality*. London: ISTE, 283–300.
[2] Devilers, R.; Bedard, Y.; Jeansoulin, R. 2005. Multidimensional Management of Geospatial Data Quality Information for its dynamic use within GIS, *Photogrammetric Engineering & Remote Sensing* 71(2): 205–215. http://dx.doi.org/10.14358/PERS.71.2.205
[3] Harvey, F. 1998. Quality Needs More Than Standards, in M. Goodchild, R. Jeansoulin (Eds.). *Data Quality in Geographic Information – From Error to Uncertainty*, pp. 192.
[4] Veregin, H. 1999. Data quality parameters, in P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind (Eds.). *Geographical Information Systems*, Vol. Principles and Technical Issues, John Wiley & Sons, Inc., 177–189.
[5] Alders, H.; Morrison, J. 1998. Spatial Data Quality for GIS, in M. Craglia and H. Onrud (Eds.). *Geographic Information Research: Trans-Atlantic Perspectives*. London/Bristol:Taylor & Francis, 463–475.
[6] Agumaya, A.; Hunter, G. J. 1997. Determining fitness for use of geographic information, *ITC Journal* 2: 109–113.
[7] Boin, A. T.; Hunter, G. J. 2008. What communicates quality to the spatial data consumer?, in A. Stein, W. Shi, W. Bijker (Eds.). *Quality aspects in spatial data mining*, 285–296. New York, NY: CRC Press.
[8] Górski, J.; A. Janowski, M. Leszczyńska. 2000. Analiza porównawcza standardów metadanych. Archiwum Fotogrametrii, *Kartografii i Teledetekcji* 2000(10): 50.1–50.5.
[9] Servigne, S.; Lesage, N.; Libourel, T. 2010. Quality components, standards, and metadata, in R. Devillers, R. Jeansoulin (Eds.). *Fundamentals of spatial data quality*, London, UK: Wiley.
[10] Wang, F.; Huang, Q. 2007. A methodology for definition and usage of spatial data quality rules, in *Geoinformatics 2007: Geospatial Information Science*, Nanjing, China, 25 May 2007 (eds) Chen J, Pu., Proceedings of SPIE, vol. 6753: D7531–D7531.
[11] Bielecka, E.; Bober, A. 2013. Reliability analysis of interpolation methods in travel time maps-the case of Warsaw, *Geodetskij vestnik* 57(2): 299–312.
[12] Burinskide, M.; Rudzkiene, V; Venckauskaite, J. 2011. Models of factors influencing the real estate price, in *Proc. of the 8th International Conference Environmental Engineering, Vilnius Lithuania 2011.* Vilnius Gediminas Technical University, 873–878.
[13] Wielebski, Ł.; Medyńska-Gulij, B. 2013. Cartographic visualization of firehydrants accessibility for the purpose of decision making, *Geodesy and Cartography* 62(2): 183–198. http://dx.doi.org/10.2478/geocart-2013-0011
[14] *The General Surveyor of Poland*, 2003, Technical Guidelines, Topographic Database – version 1, the Head Office of Geodesy and Cartography. Główny Geodeta Kraju, 2003, Wytyczne techniczne, Baza danych Topograficznych – wersja1, GUGiK.