IMPROVEMENT OF DIGITAL GEOGRAPHIC DATA QUALITY

Václav Talhofer
vaclav.talhofer@unob.cz

Šárka Hošková – Mayerová*
sarka.mayerova@unob.cz

Alois Hofmann
alois.hofmann@unob.cz
University of Defense, Faculty of Military Technology,
Brno, Czech Republic

ABSTRACT

The article deals with a spatial data quality and the decision making process. When the quality of used digital spatial data is not considered, the weak points of spatial analyses are arising. The method of geographic databases utility values is proposed. The system for utility value improvement, concerning the task to be solved is described. The expenses for data collection and data management are considered.

Keywords: digital geographic data, quality, benefit cost evaluation, decision making process

1. Introduction

The end user of digital geographic information (DGI) has to obtain not only data, but also the information about their properties. In the case of primary data this information should be given by the data producer and its content should be in accordance with e.g. the ISO Quality Standards. ISO 19113 defines data quality elements and sub-elements and then using ISO 19114 the data can be evaluated and the quality results can be reported in metadata according to ISO 19115 or in a separate quality report (Jacobsson & Giversen, 2007)

The authors considered the quality not only from the data producer´s point of view but also from the user point of view and moreover with respect to given tasks in which data are used. The general quality concept consists of various elements as technical functionality, dependability, ecology, economy, safety, etc. Not only the technical functionality is necessary to assess, but also the dependability (ability to perform as and when required), the economy (appreciation of DGI functionality and spent expenses) etc. The systems of DGI evaluation from the user´s point of view are very important and the Value Analysis Theory (VAT) (Miles, 1989) can be applied .

2. Assessment criteria for digital geoinformation quality evaluation

Using VAT the "user functions" and criteria for their evaluation were defined and used for DGI quality evaluation. The authors have derived five essential criteria from DGI review of demanded properties - *database content, database technical quality, database timeliness, area importance, user friendliness*. Their assessment gives the baseline for relatively reliable determination of each product utility value (Talhofer, Hošková, Kratochvíl, & Hofmann, 2009). Each of the criteria is mathematically assessable through independent tests.

* Corresponding author

2.1 Database content

The *database content* criterion expresses mostly compliance of the defined content and users' needs. The criterion is divided into sub-criteria. The first group includes the *real world model integrity* criterion to assess the concord of the built model and the users' requirements. It is defined as follows: $k_{11} = 100 - \alpha_{11}$, where α_{11} is a value within the 1-100 scale expressing the degree of non-conformity with the users' requirements. The other criteria group consists of *required data resolution level compliance* criteria. The criterion is divided into two sub-criteria - *geometric resolution level compliance* and *thematic resolution level compliance*. Both criteria are expressed in the form of complying objects and phenomena percentage out of the total number of all modelled objects and phenomena defined in the database concerned:

$$k_{12i} = 100\frac{n_{12i}}{n_d}, i = 1,2 .$$

- n_d is the number of all objects and phenomena defined in the database,
- n_{121} is the number of objects and phenomena in the database compliant to the users' requirements as long as the geometric resolution level concerned,
- n_{122} is the number of objects and phenomena in the database compliant to the users' requirements as long as the thematic resolution level concerned.

The total value of the *data base content* criterion may be expressed in the following equation:

$$k_1 = \left(p_{11}k_{11} + \sum_{i=1}^{2} p_{12i}k_{12i} \right) \left(p_{11} + \sum_{i=1}^{2} p_{12i} \right)^{-1} \qquad (1)$$

in which p_{1i} are sub-criterion weights. Their values are given from direct estimate or paired comparison method. (Note: The weights *p* of next criteria can be given in the same way.)

2.2 Database technical quality

The technical quality of the database is an important criterion of a strong influence on utility value and technical quality of DGI. (see (DGIWG-500, 2010), (STANAG 2215, 1989)), it is divided into five sub-criteria.

2.2.1 Transparent source data and methods used for secondary data derivation

The first part of the sub-criterion is the *level of knowledge of source information for primary data collection*. If the exact characteristics are known, the criterion value is 100, othewise the value is decreased by the percentage of the unknown or incomplete information expressed as number α_{211}. The methods and mathematic models used in secondary data derivation may considerably affect data output accuracy. Therefore the *technically correct use of secondary data derivation methods and models* make the sub-criterion other part. Similar to the previous criterion, its value equals 100 if the database designer provides complete information. If the exact information of applied methods or models is unknown, the value is decreased by the percentage of the unknown or incomplete information expressed as number α_{212}. Then the following holds: $k_{21i} = 100 - \alpha_{21i}, i = 1,2$

The k_{21} sub-criterion aggregated value is defined in the following equation:

$$k_{21} = \left(\sum_{i=1}^{2} p_{21i}k_{21i} \right) \left(\sum_{i=1}^{2} p_{21i} \right)^{-1} \qquad (2)$$

2.2.2 Positional accuracy

The sub-criterion - *positional accuracy* – assesses the accuracy of objects and phenomena locations in the given geodetic reference systems in both *horizontal* and *altitude accuracy* of the objects and phenomena. An independent test of positional accuracy proves justice or injustice of the category classification (e.g. (STANAG 2215, 1989)). Criterions k_{221} and k_{222} then evaluate the product utility as follows: $k_{22i} = 100\frac{n_{22i}}{n} + h_s, i = 1,2$, where

- n is for the total number of objects and phenomena in the used database,

- n_{22i} is for the number of objects and phenomena in the database that comply particular category horizontal or altitude accuracy, respectively,
- h_s is for selected reliability level in per cent.

Then, the k_{22} sub-criterion aggregate value is: $k_{22} = \left(\sum_{i=1}^{2} p_{22i} k_{22i} \right) \left(\sum_{i=1}^{2} p_{22i} \right)^{-1}$

2.2.3 Attribute accuracy

The product function ability is assessable from the independent test results with *attribute accuracy* k_{23} criterion as the correct (to the particular class) thematic attributes objects and phenomena percentage of all objects and phenomena in the database. The following holds:

$$k_{23} = 100 \frac{n_{23}}{n} + h_s \qquad (3)$$

- n is for the total number of all the objects and phenomena in the database,
- n_{23} is for the objects and phenomena in the database compliant to the attribute accuracy class,
- h_s is for the chosen reliability level in per cent.

2.2.4 Data base logical consistency

Database logical consistency evaluates degree of adherence to logical rules of data structure, attribution and relationships (*topological consistency*, thematic *and time consistency)*. The value of criteria k_{24i} is expressed as the percentage of the consistent objects of the all objects in the database. Then the value of k_{24} sub-criterion is calculated in the same way as for the preceding criteria, thus:

$$k_{24} = \left(\sum_{i=1}^{3} p_{24i} k_{24i} \right) \left(\sum_{i=1}^{3} p_{24i} \right)^{-1} \qquad (4)$$

2.2.5 Data completeness

Data completeness evaluates the completeness rate of all specified objects and their characteristics. *Integrity* of individual *objects* and *integrity* of their *thematic attributes* is assessed. Both the criteria are expressed as percentage of all objects and phenomena in the whole database or its part from area of interest (k_{251}, k_{252} criteria). The aggregate value of k_{25} sub-criterion is calculated as follows:

$$k_{25} = \left(\sum_{i=1}^{2} p_{25i} k_{25i} \right) \left(\sum_{i=1}^{2} p_{25i} \right)^{-1} \qquad (5)$$

So, the aggregate value of k_2 criterion to evaluate the data base quality is:

$$k_2 = \left(\sum_{i=1}^{5} p_{2i} k_{2i} \right) \left(\sum_{i=1}^{5} p_{2i} \right)^{-1} \qquad (6)$$

2.3 Database timeliness

The *database timeliness* level changes relatively fast. Its value is principally expressible as percentage of changes occurred in all the geometry, topology and/or attributes of the objects and phenomena. Nevertheless, it seems useful to assess the timeliness rate as a time function measured since the last database update. The function that expresses the overall change in the database content timeliness is a function of time and can be expressed within appropriate mathematical formula *f(T)* which expresses time obsolescence of the database content at time *T*. Then the value of criterion is $k_3 = 100 f(T)$.

2.4 Area importance

The criterion of *area importance* issues from user needs so that their processed or supported area range requirements are met. The significance of the criterion considers different importance of the same area for different users, such as military, political, economic and others. The area importance assessing criteria express the characteristics of the area and events that have been, are or will be occurring in it related to the area causing or having raised either directly or implicitly interest in the

area. When DGI is used for military purposes, the following structure of sub-criteria can be considered (e.g. geographic location of the given area, access corridors to the area of interest, amount and nature of obstacles, industrial areas, population density, location of area defence systems, etc.)

The mentioned criteria are far from complete the list and can be later amended; reduced, combined etc. Each of the criteria has own weight being mostly determined on user survey basis, such as paired comparison method. The final importance level of an area through sub-criteria assessment may be determined using the following aggregation function:

$$v_j = p_1 v_1 \sum_{i=2}^{n} p_i v_{ij} \qquad (7)$$

- v_j … overall assessment of the j^{th} square unit,
- v_{ij} … partial assessment of the j^{th} unit according to the i^{th} criterion ,
- p_i … weight of the i^{th} partial criterion,
- n … overall number of the applied partial criteria.

The criterion resultant value of area importance k_4 is then given by: $k_4 = 100 v_j$.

2.5 Data standards, independence and security

The criterion *standards, independence and security of data* means data usability in different GIS software environment, independence of data of particular software environment and, last data security system against damage or misuse. This criterion has three sub-criteria – data standards, data independence of software environment and data security against damage or misuse.

2.5.1 Data standards

The standards principally consist in the agreement of involved parties on providing data to each other in standard exchange formats to avoid troubles in the systems that support the standards. However, important for the users is whether the data are or are not provided in standard format. Therefore, the value of k_{51} criterion is $k_{51} = 0$ for disrespected the specific standard and $k_{51} = 100$ for respected the specific standard.

2.5.2 Software independent data

The data software independence means primarily the data are usable in different software environments without any modification necessary for the full utility value. The assessment of k_{52} criterion consists only in decision whether data are or are not software dependent, thus $k_{52} = 0$ for provided data dependent on data producer's software and $k_{52} = 100$ for data independent of data producer's software.

2.5.3 Data dependability, security against damage or misuse

Data dependability and security is a system of measures to prevent data from incidental or malicious damage, misuse or loss. The components of data dependability and security for production technology are excluded from this assessment. The user main data security consists of the user access to the database in time when required, user access rights to the databases, copyright system, data security while handled or transported to the users.

Each of the sub-components is evaluated with security grade within a hundred point scale. The value 100 means complete security and coefficient α is for a criterion breach deduction; the i^{th} sub-component assessment is then $k_{53i} = 100 - \alpha_{53i}$. All the sub-components have equal weights in aggregate data security, so $k_{53} = \frac{1}{n}\sum_{i=1}^{n} k_{53i}$, where n is for the number of all criterion sub-components.

The aggregate value of the criterion k_5 - standards, independence and security of data may be written as the following function:

$$k_5 = \left(\sum_{i=1}^{3} p_{5i} k_{5i} \right) \left(\sum_{i=1}^{3} p_{5i} \right)^{-1} \qquad (8)$$

2.6 General assessment of spatial data utility

The product (the whole spatial database or its part covering the given AOCR) can be assessed based on the above mentioned criteria using a suitable aggregation function F (Talhofer, Hošková, Kratochvíl, & Hofmann, 2009):

$$F = p_3k_3p_4k_4(p_1k_1 + p_2k_2 + p_5k_5) \quad (9)$$

The chosen form of the aggregation function concerns also the case when the user gets data of an area beyond his interest or redundant data; so that their use could seriously affect or even disable the DGI functions. The weight of each criterion is marked as p_i, where i = 1,…,5. The mentioned aggregation function proves the product status instantly and its utility rate. It is also applicable to experiments to find the ways to increase product utility at minimum cost increment.

2.7 Individual DGI benefit cost assessment structure

The DGI are usually developed and maintained by individual parts of the complete database, such as save units, map sheets etc. Therefore, it is necessary to assess their utility value in the above-described system within the established storing units introducing *individual benefit value*. Similarly, the individual benefit value can be applied for the selected part of master databases from the given *area of interest*.

When assessing database utility, it is useful to define *ideal quality level* at first. The ideal level is used as a *comparison standard* to express each criterion compliance level. Using the comparison standard the individual criteria compliance level and consequently aggregate utility is assessed. The compliance level of each individual criterion $u_{n,s}$ is given as follows: $u_{n,s} = \frac{k_s}{k_s^*}$, where:

- k_s is the value of s^{th} criterion compliance,
- k_s^* is the level of compliance of s^{th} criterion or its group criterion of the comparison standard.

Then the aggregate individual benefit value (*individual functionality – U_n)* of the n^{th} save unit is defined by the aggregation function of the same type as (9). Therefore:

$$U_n = p_3u_{n,3}p_4u_{n,4}(p_1u_{n,1} + p_2u_{n,2} + p_5u_{n,5}) \quad (10)$$

The individual criteria weights are identical with the weights in database utility value calculation.

Particular criteria usually consist of several sub-criteria (see (Talhofer, Hofmann, Hošková-Mayerová, & Kubíček). The authors took 20 criteria into their consideration; hence the equation for calculation the aggregate individual utility value is therefore a function of 20 variables that characterise the levels of compliance for each individual criterion.

Any modification of selected criterion has an impact on the value of U_n. Individual variables are independent one to another, so the derivation of the function can model the changed utility values or individual utility values.

$$dU = \frac{dU_n}{du_{n,i}} \quad (11)$$

where i = 1,…, 5, n = 1,…, N, and N is number of all saved units in the database.

2.8 Improved geographic service products utility using value analysis

Database functionality degree is comparable to the cost necessary for provisions—direct used material, wages, other expenses (HW, SW, amortisation, costs for co-operations, tax and social payments etc.), research and development cost, overhead cost and others. Functionality and cost imply the *relative cost efficiency (RCE)* calculated as follows:

$$RCE = \frac{F}{\sum_{i=1}^{n} E_i} \quad (12)$$

where $i = 1,\ldots, N$. Similarly to individual utility value U_n, it is possible to consider the impact of particular variables of expenses E_i on final *RCE*. The goal is to find such solution as the functionality will be maximised and the expenses will be minimize.
It is possible to find the most suitable option using *RCE*. The presented model functionality is shown in the following table (Table 1). In the initial stage, the database degree of functionality *F* is 0.5238 for one tile of Digital Land Model of the Army of The Czech Republic (DMU25). In cases 1 to 5, there are various attitudes to improve its properties – more database update (case 1), increased stored features amount (case 2), completing all missing features (case 3), completing all missing thematic properties (case 4) and completing all missing features and thematic properties (case 5). The cases 4 and 5 proved as the most functional ones. But if expenses are calculated, case 3 is the most effective output.
The described model brings no absolute solution, but it can represent a useful tool for DGI utility value assessment as well as for finding economic ways how to increase this value even under personnel or financial restrictions.

Table 1 Model of RCE calculation in a currency unit

Case	Initial	1	2	3	4	5
	T=5, a_{11}=20, n_{251} = 99, n_{252} = 50	T=1, a_{11}=20, n_{251} = 99, n_{252} = 50, difficulty class 3	T=1, a_{11}=15, n_{251} = 99, n_{252} = 50, difficulty class 4	T=1, a_{11}=20, n_{251} = 100, n_{252} = 50, difficulty class 3	T=1, a_{11}=20, n_{251} = 99, n_{252} = 100, difficulty class 4	T=1, a_{11}=20, n_{251} = 100, n_{252} = 100, difficulty class 4
F	0.5238	0.6734	0.6815	0.6737	0.6856	0.6859
RCE		2.8878	2.4965	2.8889	2.5116	2.5126
Δ RCE			0.3913	-0.0011	0.3762	0.3752

3. Conclusion

The presented process of VAT utilization of the DGI quality assessment is applicable to evaluation of present products as well as planned products. When this model is used for a present product, it is possible to optimize its characteristics. In the case of a planned product, it is possible to assess various variants.

Acknowledgement

The theory and results presented above were developed within the project "The Evaluation Of Integrated Digital Spatial Data Reliability" (project No.: 205/09/1198) funded by the Czech Science Foundation.

REFERENCES

DGIWG-500. (2010). *Implementation Guide to the DGIWG Feature Data Dictionary (DFDD)* (2.2.2 - 19 July 2010 ed.). DGIWG.
Jacobsson, A., & Giversen, J. (2007). *Eurogeographics*. Retrieved 2009, from http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf
Miles, L. D. (1989). *Techniques Of Value Analysis Engeneering* (3rd ed.). USA: Eleanor Miles Walker.
STANAG 2215. (1989). *Evaluation of Land Maps, Aeronautical Charts and Digital Topographic Data* (5 ed.). NATO Military Agency for Standardization (MAS).
Talhofer, V., & Hofmann, A. (2009). Possibilities of evaluation of digital geographic data quality and reliability. *24ht International Cartographic Conference, The World's Geo-Spatial Solutions*. Santiago de Chilie: ICA/ACI.
Talhofer, V., Hofmann, A., Hošková-Mayerová, Š., & Kubíček, P. Spatial data quality and decision making process. *Intenational Cartographic Conference ICC 2011* (p. 14 pp). Paris: ICA.
Talhofer, V., Hošková, Š., Kratochvíl, V., & Hofmann, A. (2009). Geospatial Data Quality. *ICMT'09 - International conference on military technologies 2009* (pp. 570-578). Brno: Univerzita obrany.