POSSIBILITIES OF EVALUATION OF DIGITAL GEOGRAPHIC DATA QUALITY AND RELIABILITY

Václav Talhofer[1]
Alois Hofmann[2]
University of Defense, Faculty of Military Technology, Department of Military Geography and Meteorology,
Kounicova 65, 662 10 Brno, Czech Republic

Abstract

The digital geographic data and the digital geographic information (DGI) have got many users from different branches and they use the data for many various tasks including the support of military and non military activities of armed forces. The properties of used data can have an important influence on final results of these tasks and therefore the data provider has to inform users about them. The system of metadata built according to ISO standards is very useful for the essential information transfer, but the system could be added by the user point of view because of the frequent different view of the data provider and data user.

According to the Value Analysis Theory user functions and criterions for their evaluation can be defined for user point of view appreciation. The mathematical formulas can be applied for the fulfillment of criteria level determination and after that the final level of complex function of user worth of the product can be evaluated. The result can be compared with expenses necessary for the achievement of demanded level of user value. The different possibilities can be considered, e.g. whether more detailed data will carry corresponding and expecting higher level of user value and therefore more quality decision can be made.

Keywords: digital geoinformation, evaluation, benefit cost evaluation, value analysis

1. Introduction

The end user of DGI has to obtain not only own 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 quality concept is described very deeply in this document and it is added by several examples from different countries in Europe.

[1] COL doc. Ing. Václav Talhofer, CSc., E-Mail: vaclav.talhofer@unob.cz
[2] Ing. Alois Hofmann, CSc., E-Mail: alois.hofmann@unob.cz

But we can look at the quality not only from the data producer´s point of view but also from the user and given tasks in which data are used. The general quality concept consists of various elements (see ***Fig. 1***)

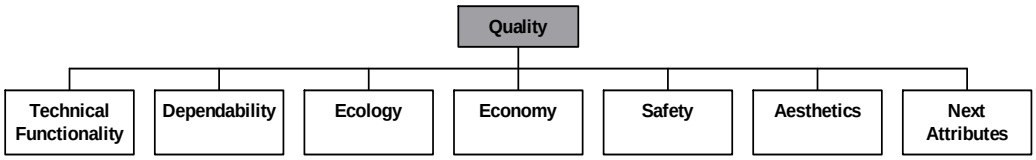

Fig. 1 The general elements of the quality

Not only the technical functionality is necessary to assess, but some other elements are useful to add into the DGI quality evaluation. The dependability (ability to perform as and when required), the economy (appreciation of DGI functionality and spent expenses) etc. are very important too.

The different organizations can participate in the data collection for one project, especially in the case of international projects (Multinational Geospatial Co-production Program – MGCP is examples of that). Used technologies for the data collection don´t have to be the same and despite of accepted standards the resulting data quality doesn´t have to be the same in the whole area of given project. But the user should be informed about the whole project or about that parts which uses in a given time on a given place.

The systems of DGI evaluation from the user´s point of view are very important and the Value Analysis Theory (VAT) ((Miles, 1989), (Crow, 2002)) can be applied. According to VAT user functions and criteria for their evaluation can be defined and used for DGI quality evaluation (Talhofer, 2004).

2. Functions of Digital Geographic Information

DGI is used to find solutions of a number of projects of various natures. The certain level of generality allows defining common features of all actions and thus to make a list of functions desired for DGI, as follows:

1. *Information function* which expresses DGI's ability of fast and reliable provision of information on position and required properties of geographic objects and phenomena in the area of interest.
2. *Model function* which expresses DGI's applicability as a model for derivation of geometric, topological or other relations.
3. *Source function for mathematical modeling, designing and planning* which is applicable to the cases that DGI use to make the future action intent or to design some work to be done.
4. *Automation function* which is evident e.g. in implementation process of control of designed and planned projects or of actual activities (e.g. moving objects monitoring).
5. *Illustration function* which expresses DGI's ability to illustrate a situation.
6. *Source function for derivation* which means an ability to derive other types of DGI from original or to be a source for maps making and for other cartographic purposes.

The above mentioned functions are uneasy to apply them directly for users' demand expression. It means it is impossible to determine their level of satisfaction. However, to carry out the functions as listed, DGI needs certain properties. Then, the assessment of those properties (criteria) shall replace the assessment of function level of satisfaction and thus it shall be determinable what level the criteria meet the defined standards, regulations, etc.

3. Function conditioned properties of DGI, assessment criteria

Five essential criteria were derived from DGI demanded properties review. Their assessment gives the baseline for relatively reliable determination of each product utility value:

1. *Database content* expresses mostly compliance of its definition and users' demands, i.e. concord of the "real modeled world" and its model represented by objects and phenomena stored in the database.
2. *Database technical quali*ty defines the technical quality of stored data.
3. *Database timeliness* tells how much are the entire database or its parts updated.
4. *Area importance* is determined by users' needs so that it meets the requirements of processed or supported area range.
5. *User friendliness.* This criterion defines data usability in various software environment types of GIS nature reflected mostly in compliance to standard principles. This criterion further reviews data dependability, independence and security.

Each of the criteria is mathematically assessable through independent tests.

3.1 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, and is defined by the following equation:

$$k_{11} = 100 - \alpha_{11} \qquad (1)$$

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 modeled objects and phenomena defined in the database concerned:

$$k_{12i} = 100\frac{n_{12i}}{n_d}, i = 1,2 \qquad (2)$$

In which

- 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:

$$k_1 = \left(p_{11}k_{11} + \sum_{i=1}^{2} p_{12i}k_{12i} \right) \left(p_{11} + \sum_{i=1}^{2} p_{12i} \right)^{-1} \quad (3)$$

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.)

3.2 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. Many publications from the geoinformation branch involve the data quality definition issue (e.g. (DFDD, 2008), (STANAG 2215, 1989)). This criterion is divided into five sub-criteria. The total technical quality assessment needs, therefore assessing of all individual elements.

3.2.1 Transparent source data and methods used for secondary data derivation

The first part of the sub-criterion is the *precise knowledge of source information for primary data collection*. Provided that the database designers know exactly the used sources characteristics this criterion value equals 100. If the exact characteristics are known only particularly, 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 and administrator provide 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 applies:

$$k_{21i} = 100 - \alpha_{21i}, i = 1,2 \quad (4)$$

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

$$k_{21} = \left(\sum_{i=1}^{2} p_{21i}k_{21i} \right) \left(\sum_{i=1}^{2} p_{21i} \right)^{-1} \quad (5)$$

3.2.2 Positional accuracy

The second sub-criterion - *positional accuracy* – is to assess 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)) and from this point of view, and then k_{221} and k_{222} criterions evaluate the product utility as in the following equation:

$$k_{22i} = 100\frac{n_{22i}}{n} + h_s, i = 1,2 \quad (6)$$

in which

- n is for the total number of objects and phenomena in the 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:

$$k_{22} = \left(\sum_{i=1}^{2} p_{22i}k_{22i}\right)\left(\sum_{i=1}^{2} p_{22i}\right)^{-1} \quad (7)$$

3.2.3 Attribute accuracy

The third sub-criterion is *attribute accuracy*. The product function ability is assessable from the independent test results with 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 applies:

$$k_{23} = 100\frac{n_{23}}{n} + h_s \quad (8)$$

in which

- 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.

3.2.4 Data base logical consistency

The fourth sub-criterion is the *database logical consistency*. This criterion evaluates degree of adherence to logical rules of data structure, attribution and relationships. The evaluated features include primarily *topological consistency* for the most applications basic qualification, then *thematic (domain) consistency* and *time consistency*. The level of *topological, thematic and time consistencies* criteria k_{241}, k_{242} and k_{243} is expressed as the percentage of the consistent objects of the all objects in the database. Independent tests are useful to use for criteria evaluation. The aggregate value of k_{24} sub-criterion is calculated in the same way as for the preceding criteria, thus:

$$k_{24} = \left(\sum_{i=1}^{3} p_{24i}k_{24i}\right)\left(\sum_{i=1}^{3} p_{24i}\right)^{-1} \quad (9)$$

3.2.5 Data completeness

Data completeness is the last sub-criterion to evaluate completeness rate of all specified objects and phenomena and their characteristics. However, a number of objects may enter the database without thematic attributes included. This is a quite frequent practice the users meet. Therefore, it is useful to assess *integrity* of individual *objects and phenomena* and *integrity* of their *thematic attributes*. Both the criteria are evaluated in per cent of all objects and phenomena in the whole database or its part from area of

interest - AOI (k_{251}, k_{252} criteria). The aggregate value of k_{25} sub-criterion is calculated with the same equation as the previous ones:

$$k_{25} = \left(\sum_{i=1}^{2} p_{25i} k_{25i} \right) \left(\sum_{i=1}^{2} p_{25i} \right)^{-1} \quad (10)$$

The aggregate value of k_2 criterion to evaluate the data base quality is calculated with the following equation:

$$k_2 = \left(\sum_{i=1}^{5} p_{2i} k_{2i} \right) \left(\sum_{i=1}^{5} p_{2i} \right)^{-1} \quad (11)$$

3.3 Database timeliness

The *database timeliness* level changes relatively fast in dependence on various factors. 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*. The assessment of database timeliness for value analysis expressed with k_3 coefficient is applicable to the following equation:

$$k_3 = 100 f(T) \quad (12)$$

3.4 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:

1. Geographic location of the evaluated area
2. Access corridors to the AOI
3. Amount and nature of obstacles
4. Industrial centers
5. Population density
6. Garrison deployment and size
7. Area defense systems/equipment deployment

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:

$$v_j = p_1 v_1 \sum_{i=2}^{n} p_i v_{ij} \qquad (13)$$

in which

- 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 calculated using the following equation:

$$k_4 = 100 v_j \qquad (14)$$

3.5 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 divides also into three sub-criteria – data standards, data independence of software environment and data security against damage or misuse.

3.5.1 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.

3.5.2 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.

3.5.3 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 following:

1. User access to the database in time when required
2. User access rights to the databases
3. Copyright system
4. Data security while handled or transported to the users, mostly via communication line

Each of the sub-components is evaluated with security grade within a hundred point scale, in which the value 100 means complete security and coefficient α is for a criterion breach deduction; the i^{th} sub-component assessment is then as follows:

$$k_{53i} = 100 - \alpha_{53i} \quad (15)$$

Provided that all the sub-components have equal weight in aggregate data security, this is expressed as:

$$k_{53} = \frac{1}{n}\sum_{i=1}^{n} k_{53i} \quad (16)$$

in which 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:

$$k_5 = \left(\sum_{i=1}^{3} p_{5i}k_{5i}\right)\left(\sum_{i=1}^{3} p_{5i}\right)^{-1} \quad (17)$$

3.6 General assessment of digital geoinformation utility value

The whole database or its used part can be assessed based on the above mentioned criteria using a suitable *aggregation function.* Degree of compliance of mentioned function can be express in the form:

$$F = p_3\, k_3 p_4 k_4 (p_1 k_1 + p_2 k_2 + p_5 k_5) \quad (18)$$

The chosen form of the aggregation function concerns also the case the user gets data on an area beyond his AOI or data obsolete so that their use could seriously affect or even disable the DGI functions. The weights of each criterion p_i are determined as the sub-criteria weights with the expert estimate methods and should be modified according to given user specification or given task. The aggregation function proves the product status at the questioned instant and its utility rate. It is applicable also to experiments to find the ways of how to increase product utility at minimum cost increment.

4. Individual DGI benefit cost assessment structure

The DGI databases are usually divided into parts of the complete database, such as tiles, map sheets etc. Therefore, it is possible to assess their utility value in the above described system within the established loading units introducing *individual benefit cost value (IBCV)*. When assessing database utility, it is useful to define *ideal quality level* at first and uses it a *comparison standard* to express each criterion compliance level. Using the comparison standard the individual criteria compliance level and consequently aggregate utility may be assessed. *Individual criteria compliance level* general equation is as follows:

$$u_s^x = k_s\left(k_s^*\right)^{-1} \quad (19)$$

in which

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

The aggregate IBCV (*individual functionality)* of the *x-th* measurement unit is defined by the aggregation function of the same type as (18). Therefore:

$$U^x = {}^\circ F^x = p_3 u_3^x \, p_4 u_4^x (p_1 u_1^x + p_2 u_2^x + p_5 u_5^x) \qquad (20)$$

The individual criteria weights are identical with the weights in database utility value calculation (18). The equation to calculate the aggregate individual utility value is therefore a function of 29 variables that characterize the levels of compliance for the individual criteria. Provided individual variables are independent of each other, the derivation of the stated function can model the changed utility values or individual utility values.

$$dF^x = \frac{dU^x}{du_i} \qquad (21)$$

The principal criteria compliance levels are functions of more variables, though. Determination of du_i value is thus feasible in two ways regarding the desired information structure. When assessing *individual variables effects* on the aggregate individual utility value while the other variables keep constant values, it is necessary to derive *U* function as follows:

$$dF^x = \frac{dU^x}{du_i^x} \frac{du_i^x}{dx} \qquad (22)$$

in which *x* is for one of the 29 mentioned variables.

However, practically multiple factors may change at the same time. An example would be the database technical quality changes all its parameters – secondary data derivation methods improve location and attribute accuracy and data integrity increase and on top of it data are stored in a geodatabase accessible to all authorized users and handled well for all topological, thematic and time relations. In such a case it is suitable to define du_i value as a total differential of all variables describing the modified factors.

5. Value improvement process

The geodatabase functionality degree is comparable to the expenses N_i for DGI provisions such as direct material, direct wages, other direct and indirect expenses (HW, SW, amortization, tax and social payments, research and development cost etc.). Functionality and cost imply *relative cost efficiency (RCE)* calculated as follows:

$$RCE = {}^\circ F \left(\sum_{i=1}^{n} N_i \right)^{-1} \qquad (23)$$

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 $^\circ F$ 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

6. 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.

7. Bibliografie

Crow, K. (2002). *Value analysis and function analysis system technique.* Retrieved 08 2009, from DRM Associates: http://www.npd-solutions.com/va.html

DFDD. (2008). *Implementation Guide to the DGIWG Feature Data Dictionary.* Defence Geospatial Information Working Group (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. (2004). Digital Geographic Data: Potential Evaluation. *AGILE 2004, 7th Conference on Geographic Information Science, Conference proceedings* (pp. 675 - 686). Heraclion, Crete, Greece: AGILE.