Modern techniques for evaluation of spatial data quality

Lect. PhD. DROJ Gabriela - gaby@oradea.ro
Lect. PhD. stud. eng. SUBA Stefan –subast@yahoo.com
Lect. PhD. stud. BUDA Aurelian - budaaurelian@yahoo.com

Abstract: *For centuries ago cartographers, geographers, surveyors and geodesists have been involved in collection, storage, analysis and visualization of spatial data. The collections of spatial data are subject to random, systematic and gross errors. In the last decade GIS became the main tool for analyzing spatial data in many real life activities. The quality of the geographical information systems is dependent on the quality of the stored and queried data. In comparison with the non-geographic information systems, in GIS applications, the quality of data has a greater influence on the results. Despite the arguments in favor of explicit treatment of spatial data quality, usually Cadastre and GIS practice continues to proceed as if data were perfect. The objectives of this study are to identify different techniques for evaluation of spatial data quality.*

Keywords: *GIS, Spatial Data Quality, cadastre*

1. Introduction

For centuries ago cartographers, geographers, surveyors and geodesists have been involved in collection, storage, analysis and visualization of spatial data. The collections of spatial data are subject to random, systematic and gross errors.

In the last decade GIS became the main tool for analyzing spatial data in many real life activities. The quality of the geographical information systems is dependent on the quality of the stored and queried data. In comparison with the non-geographic information systems, in GIS applications, the quality of data has a greater influence on the results.

The five main reasons for the main concerns about spatial data quality issues were identified as:

a) There is an increasing availability, exchange and use of spatial data;
b) There is a growing group of users less aware of spatial data quality;
c) GIS enables the use of spatial data in all sorts of applications, regardless of the data quality;
d) Current GIS offers hardly any tools for handling spatial quality;
e) There is an increasing distance between those who use the spatial data (the end users) and those who are creating the spatial databases [van Oort].

Despite the arguments in favor of explicit treatment of spatial data quality, usually Cadastre and GIS practice continues to proceed as if data were perfect. Data are used uncritically without consideration of the error contained and this can lead to erroneous results, misleading information, unwise environmental decisions and increased costs.

The objectives of this study are to identify different techniques for evaluation of spatial data quality.

2. The quality of spatial data

The quality of spatial data is defined by a set of characteristic classified in [Dro09]:

- *Dependability characteristics* refer to the time related aspects of quality: availability, temporal correctness. The temporal correctness is given by the moment of collection of the data and by the level of updating.
- *Integrity characteristics* represent the applicability of information, the number of collected data and their properties: completeness, consistency, correctness.
- *Accuracy characteristics* refer the positional accuracy and attribute accuracy.

The requirements for the quality of the spatial data are depending on the specificity of the informational system. For example the quality requirements for the cadastres data are higher than for the environmental application. On the other hand the quality requirements for cadastre data are usually imposed by local and regional standards.

ISO 19113 provides a classification scheme for these errors. They are categorized into different elements and sub elements depending on the nature of the error. Regarding the ISO 19113 standard the data quality aspects are [ISO]:

- Completeness – presence and absence of features and their attributes and relationships:
  - Commission – excess data presents in a dataset,
  - Omission - data absent from dataset,
- Logical consistency – degree of adherence to logical rules of data structure, attribution and relationships:
  - Conceptual consistency - adherence to rules of conceptual schema,
  - Domain consistency – adherence of values to the value domain,
  - Formal consistency – degree of which data is stored in accordance with physical structure of dataset,
  - Topological consistency – correctness of explicitly encoded topological characteristics of a dataset
- Positional accuracy – accuracy of the position of feature
  - Absolute or external accuracy - closeness of reported coordinate values to values accepted as or being true,
  - Relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true,
  - Gridded data position accuracy - closeness of gridded data position values to values accepted as or being true,
- Temporal accuracy – accuracy of temporal attributes and temporal relationships of features
  - Accuracy of a time measurement – correctness of a temporal references of an item (reporting of error in time measurement)
  - Temporal consistency – correctness of ordered events or sequences, if reported,
  - Temporal validity – validity of data with respect to time
- Thematic accuracy – accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships
  - Classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (ground truth or reference dataset),
  - Non-quantitative attribute correctness – correctness of non-quantitative attributes,
  - Quantitative attribute accuracy – accuracy of quantitative attributes.

3. Cadastral issues

Cadastral survey or cadastral map is a comprehensive register of the metes-and-bounds real property of a country. A cadastre commonly includes details of the ownership, the precise location, the dimensions (and area), the cultivations if rural and the value of individual parcels of land. Cadastral information are important in both aspects the legal and spatial part of the information, however the legal issues is mostly dominating. Cadastre in European Union does not reply to a unique model, there are used various types of cadastre inventory systems based on different patterns and functions. The purpose of European Union is not to create a new cadastral system but to this variety without reducing efficiency of future common developments.

European cadastre's can be categorized as: [Jak09]

- **Positional-accurate cadastre** in positional-accurate type cadastres the attribute (ownership records) and the location are both up-to-date and accurate. These can be found in Austria, the Netherlands and Finland (urban areas).
- **Index map cadastre** In index map type cadastres, the actual representation of the property units can be found either from legal documents or in the field but the ownership records are accurate and up-to-date. Index maps present the topological relations between the cadastral units. Examples of these cadastres can be found in Sweden and Finland (outside urban areas).
- **Mosaic cadastre** In mosaic type cadastres there in no legal requirement to register the ownership and therefore the cadastre often covers only part of the country. The Cadastre contains information of individual ownerships, but there is no direct relationship with neighboring parcels. Combinations of different title maps may be compiled or reproduced but the topological relations remain undefined. These cadastres exist in Great Britain.

Regarding the requirements of the new cadastral law (7/1996), the Romanian cadastre will be in the category of positional accurate cadastre, but nowadays the are several parallel cadastral systems in use in Romania. The cadastral parcel is an important component not only in national but also in European spatial data infrastructures. In the context of global economy is necessary to create effective infrastructures and promote efficiency, therefore the systems of spatial data had to be interoperable.

4. Standards of spatial data

To achieve interoperability of geographic information is necessary to establish a set of minimum standards and policies. This standards and policies must define at three levels[Dro09]:

- *Institutional level* represented by organizational rules like access rules, data protection and copyright,
- *Technical level* consists in all the hardware, software and communication protocols compatibility,
- *Semantic level* - data standards, the set of public data and process standards.

In this paper we will focus on the semantic level. **Data standards** are semantic definitions that are structured in a model. They describe the minimum requirements of objects, features or items that are (will be) collected, automated, or affected by process. **Process standards** also referred as service standards describe how the procedures to follow, methodologies to apply, procedures to present information, or business process rules to be followed to implement other standards. Process standards are used: (1) to establish a threshold for minimally acceptable data, (2) to determine the best data for an application, or (3) to promote interoperability and broad based use of data.

Current standards in geographic information may be classified as industry standards also called de-facto standards and official standards, national, regional, or international standards. The de- facto standards are generated by industry, in the case of spatial data this is the Open GIS Consortium (Open Geospatial). The official standards can be also divided in international standards (ISO), regional standards (European standards INSPIRE) and national standards.

These standards specifies all the methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations.

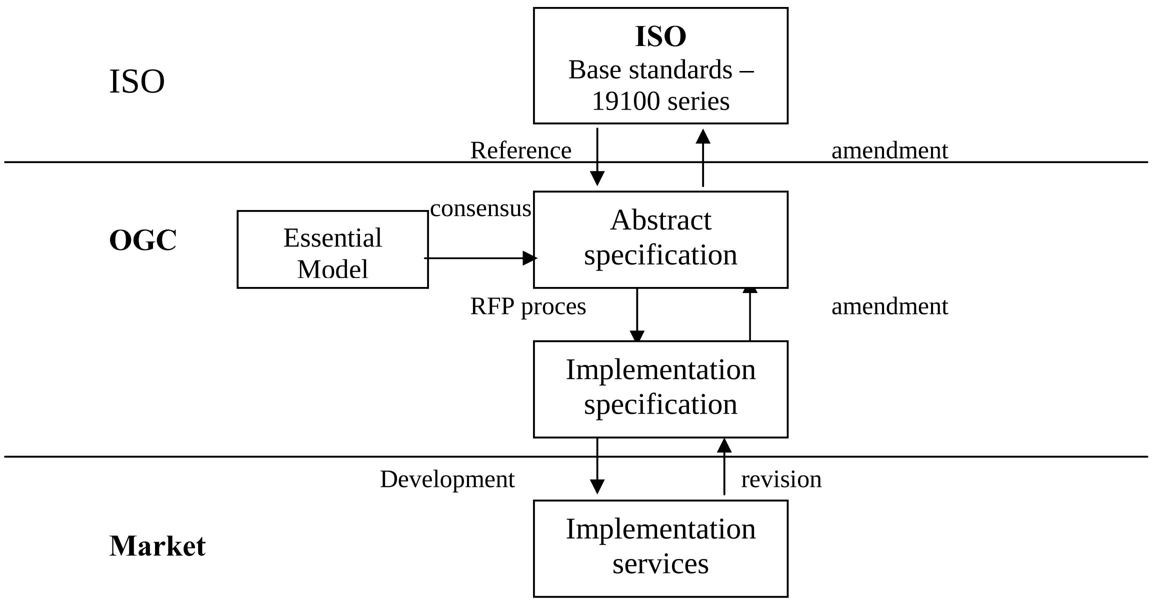

Figure 1 - Relationship between ISO and OGC standards [Oos02]

Being international standards ISO is respected by the entire official standards or de – facto. The standardization in the field of digital geographic information is done by TC211 committee. Their work aims to establish a structured set of standards known as the ISO 19100 series of geographic information standards. These standards are meant to enable geospatial datasets to interact between different data models and different applications. The most important quality-related standards, for spatial data, in the ISO 19100 families are:

- ISO 19113 Geographic Information – Quality principles
- ISO 19114 Geographic Information – Quality evaluation procedures
- ISO 19138 Geographic Information – Data quality measures
- ISO 19115 Geographic Information – Metadata
- ISO 19115 Geographic Information – Metadata –Part 2 : Extensions for imagery and gridded data
- ISO 19122 Geographic Information – Qualifications and certification of personnel
- ISO 19127 Geographic Information – Geodetic codes and parameters
- ISO 19131 Geographic Information – Data product specifications
- ISO 19139 Geographic Information – Metadata – XML schema implementation

The must important industrial standards for spatial data specification are the OpenGIS. **OpenGIS ® Standards** and Specifications are technical documents that detail interfaces or encodings. Software developers use these documents to build support for the interfaces or encodings into their products and services [Ogis].

There is a close link between the ISO and OpenGIS standards. These relations are presented in figure 1.[Oos02]:

Each country, based on ISO 19100 series, OpenGIS Standards and Specifications and national land management laws, tries to standardize the spatial data in order to establish a National Spatial Information Data. For European countries this NSDIs must also fulfill the regulation of INSPIRE.

INSPIRE (Infrastructure for Spatial Information in Europe) is meant to create a legal framework in order to establish and operate an Infrastructure for Spatial Information in Europe, for the purpose of formulating, implementing, monitoring and evaluating Community policies at all levels and providing public information.

5. Techniques for evaluation of spatial data quality

The international ISO standards provides principles not only for the principles how geospatial data can be described but also offer guidance on assessing the quality of actual data sets and what should be the quality evaluation procedures. The procedures for evaluation of data quality according to the ISO 19114 standard it must be done in five steps[Jak09, Joo06]:

- Identifying the data quality scope: elements and subelements,
- Identifying the data quality measure
- Selecting the evaluation method,
- Determine data quality results,
- Determine conformance.

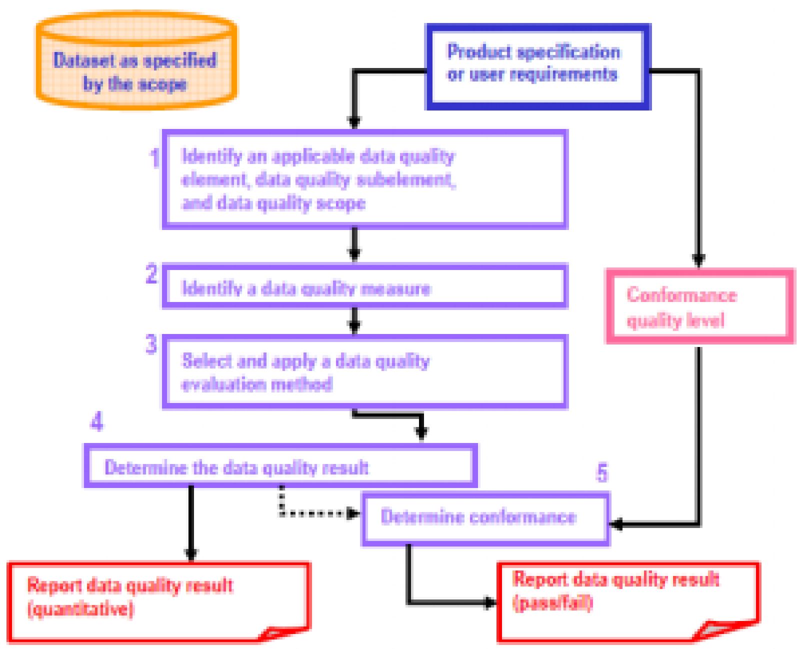

Figure 2 - Five steps for quality evaluation according ISO 19114

The five steps of quality evaluating procedures and the relationship between them are presented in the upper picture.

The quality evaluation information shall be reported as metadata. A separate quality evaluation report is required when metadata result is only pass/fail or when aggregate quality results are generated.

The standard specifies the fields to be filled in when reporting on assessment as a quality evaluation report (Annex I in the standard)[Jak09]:

- Identification of reporting document
- Scope observed
- Measure used (formula, resulting values, result unit, reliability, reliability unit)
- Confidence in conformance test (confidence value, confidence unit, documents explaining the method)
- Type of quality evaluation method used (direct external etc., inspection strategy applied)
- Description of quality method used (basic assumptions, processing algorithms, definition of parameters, parameter values for the specific test, parameter units),
- Possible aggregation of results (unit for aggregated values, resulting values, statistics used for aggregation, computation date, pointer to aggregation report)

The evaluation of spatial data quality can be done by using two methods: by direct observation, with visual comparisons of the resulted models, and by using statistical parameters.

The visual comparisons are very difficult to be done and impossible to be quantified. In the case of quality evaluation made through visual comparison is required to have images of the both models, in order to model a real life situation and the generated model.

In order to quantify the uncertainty of the spatial data, it is necessary to use statistical methods. To decide the quality of the data by statistical analysis, the comparison of the known data with the resulted data is required. To analyze the pattern of deviation between two sets of data, conventional ways are to yield statistical expressions of the accuracy, such as the root mean square error, standard deviation, mean, variance and coefficient of variation [Dro08].

The quality requirements for cadastre data are also dependent of the type of inventory system used. In the following table we present the role of quality in the different type of cadastre. [Jak06, Jak09]

Cadastre	**Type Ownership data**	**Location data**	**Potential quality issues**
Mosaic	Ownership records	Ownership maps (general boundaries)	Reliability No other quality information available
Index	Ownership as attribute information	Uniform index map	Positional Accuracy Topological consistency Process management
Positional-accurate	Ownership as attribute information	Accurate location of boundary points	Logical consistency Security issues Process management

The ISO 19138 specifies the data quality measures; it defines a set of data quality measures that can be used when reporting data quality for the sub-elements in ISO 19113. In the table below we presented the most important quality measures for cadastral issues selected from ISO 19138[Jak06, Jak09].

Data quality element	**Data quality sub-element**	**Data quality measure**	**Data quality basic measure**
Completeness	Commission	Number of excess items	Error count
Completeness	Omission	Number of missing items	Error count
Logical consistency	Topological consistency	Number of faulty point-curve connections	Error count
Logical consistency	Topological consistency	Number of missing point-connections	Error count
Positional accuracy	Absolute or external accuracy	Covariance matrix	
Positional accuracy	Absolute or external accuracy	RMSE or standard deviation, where the true value is not estimated from the observation but known a priori	
Temporal accuracy	Absolute of time measurement	Mean value of date attributes	Difference between date in database and date
Temporal accuracy	Temporal validity	Number of items not in conformance with their value	Error count
Thematic accuracy	Classification correctness	Number of incorrectly classified features	Error count
Thematic accuracy	Classification correctness	Misclassification matrix	

6. Conclusion

In this paper we analyzed the main issues regarding spatial data quality in general and in cadastre. We studied a classification of the characteristics which represent the quality of spatial data and the main elements and subelements of spatial data quality as it is regulated by ISO standards. Also we present a brief introduction in the guidelines established by different standardization organization and the relationships between them.

The main conclusion which is pointed up by the studies is the importance of standards in data quality assessment.

The second conclusion consists in the usage of spatial data standardization as a technique for evaluation of spatial data quality.

The last conclusion is that, there are two keys to the improvement of data quality – they are prevention and correction. Error prevention is considered to be far superior to error detection, since detection is often costly and can never guarantee to be 100% successful, and the best way of error prevention is to follow all the regulation regarding specification, structure, data sets, and implementation and evaluation procedures.

References:

[Dro08]Droj, G - Quantifying the Uncertainty of Spatial Data, ZAC, Cluj, 2008
[Dro09] Droj, G, Suba, S. - Role of Data Quality in GIS Decision Support Tools, Szekesfehervar GIS Open 2009

[Jak09]Jakobsson, A. Giversen, J. - Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies, www.eurogeographics.org, visited in march 2009

[Jak06]Jakobsson, A. Giversen, J. - On the Future of Topographic Base Information Management in Finland and Europe, Dissertation for degree of Doctor in science, 2006,

[Joo06]Joos G. – Data Quality Standards, XXIII FIG Congress , 2006

[Ogis99] OpenGIS® - Simple Features Specification For SQL Revision 1.1, Open GIS Consortium 1999

[Ogis05] OpenGIS® - Implementation Specification for Geographic Information - Simple feature access - Part 2: SQL option, Open Geospatial Consortium 2005

[Ost02] Østensenstensen, O; Smits P. - ISO/TC211: standardisation of Geographic Information and Geo-Informatics, Canada International Geoscience and Remote Sensing International Symposium 2002

[van Oort] van Oort, P. -Spatial data quality: from description to application, NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission Delft, December 2005.