Quality Evaluation of Spatial Data in Raster Format

JITKA KOMARKOVA, PAVEL SEDLAK, PETR SRAMEK, JAN PITTERLE
Faculty of Economics and Administration
University of Pardubice
Studentska 95, 532 10 Pardubice
CZECH REPUBLIC
jitka.komarkova@upce.cz, pavel.sedlak@upce.cz

Abstract: - A fast access to all relevant information is today a necessary precondition for decision-making process. It is supposed that information provided by an information system is at a sufficient quality level. To ensure a good quality of output information, only data of a good quality must be used. Therefore, many standards and data quality models have been developed. Later on, attention was focused on spatial data too. Spatial data have several specific properties so they cannot be evaluated directly according to the general quality models. The paper brings a short overview of spatial data quality evaluation and than it focuses on raster data. The main aim of the paper is to propose a set of quality characteristics and parameters for raster data within the framework of existing ISO standards. The proposed set is used to evaluate example data, which were chosen for solution of the given task. Results obtained using good and inappropriate data are shown too.

Key-Words: - Data quality, Spatial data, Raster data, Geodata, Raster data quality, Quality parameters,

1 Introduction
Data quality significantly influences results of decision making process. Information used to support decision-making process is a result of processing data by an information system. So, data quality has been widely recognized as an important issue connected to information systems and their quality and usefulness. It is one of the reasons why many standards of data quality and several directives, laws exist, e.g.:

- Data Quality Act (USA, Public Law 106-554)
- ISO 8000 – Data Quality, parts 100, 102, 110, 120, 130, 140

It is quite difficult to define the term quality. According to the ISO 9000 quality is [1]: „the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.” SDMX (Statistical Data and Metadata Exchange; the BIS, ECB, EUROSTAT, IMF, OECD, UN, and the World Bank common initiative) defines quality as [2]: “the degree to which a set of inherent characteristics fulfils requirements.” In general, data quality is defined as a set of characteristics [3] or as a multifaceted concept with several different dimensions [5]. All authors mention context of use of data and information as an important factor which significantly influences requirements on data quality. Customers (users of data and information) and their needs must be taken into account too. So, data should fit to their use.

One example of data quality model consisting of categories and dimensions [4]:

- Intrinsic quality: accuracy, objectivity, believability, reputation
- Accessibility: accessibility, access security
- Contextual quality: relevancy, value-added, timeliness, completeness, amount of information
- Representational quality: interpretability, ease of understanding, concise representation, consistent representation, ease of manipulation.

SDMX defines the following data quality dimensions [2]:

- Relevance
- Accuracy
- Timeliness
- Punctuality
- Accessibility
- Clarity / interpretability
- Comparability
- Coherence
- Integrity
Credibility
Methodological soundness.

For some decision-making processes special data are required. Many activities of mankind happen somewhere, i.e. they can be located in a space.

Importance of support of spatial decision-making process has been recognized for a long time [6], [7] and today it is obvious to use spatial decision support systems (SDSS) to support decision-making. Geographic information systems (GIS) can be used to manage knowledge [8], GIS-based SDSS for reclamation on lands contaminated from coal mine waste [9], transportation network design [10], land use allocation [11], planning urban infrastructure [12] or evaluation of land consolidation projects [13].

The paper discusses several quality models. Within ISO standards framework it proposes a suitable set of parameters and their values with focus on raster data suitable for selected types of spatial analyses. In the paper the terms “data” and “information” will be used similarly because it is not necessary to distinguish between them.

2 Problems with Spatial Data Quality Evaluation
Information systems should support decision-making process by providing good information. Good information can be provided only in the case that data of a good quality are available. As it was mentioned before, data quality must be evaluated according to users’ expectations (both implicit and explicit) and according to the context of use of data and resulting information.

Because spatial data are used to support many important decision-making processes, their quality must be at an appropriate level. This fact resulted in several spatial data quality directives and standards:
- ISO 19100 Series – Geographic Information Quality Standards; adopted by The European Committee for Standardization (CEN)

Data quality concept according to ISO 19113 is shown in the Fig. 1.

Because of the special nature of spatial data general data quality models must be fitted to the needs of spatial data quality evaluation. Spatial data describe reality from three main points of view:
- Geometric (i.e. location, size, shape, topology)
- Thematic (attributes, non-spatial properties of features)
- Temporal (changes in time).

All these issues must be evaluated by data quality model and its characteristics.

It was realized in the remote past that electronic spatial data quality (suitable for GIS and SDSS) should be evaluated in another way than analogue spatial data quality (i.e. maps) [15].

Besides existence of ISO standards, European project EuroRoadS [16] proposed the following quality model for the purpose of collecting European road data – see Fig. 2.
The main reason for this proposal was to create a quality model with just one quality characteristic for each quality phenomena and to clearly cover all three parts of feature description (i.e. geometric, thematic and temporal) [16]. The final result is shown in the Fig. 3.

Fig. 3 - Quality elements of ISO 19113 and quality characteristics in EuroRoadS quality model [16]

The above described approach of the EuroRoadS project is very inspiring because of the clearness of the proposed quality model. There is one significant property of the model and its parameters – it is focused on vector data and data about the road transportation only. It does not deeply deal with raster data which belong to the important types of spatial data suitable for many spatial analyses.

Aim of this paper is to fill this gap and propose a suitable set of quality characteristics and parameters for raster data within the framework of existing ISO standards.

3 Quality Characteristics and Parameters for Raster Data

The following proposal of data quality parameters is focused on raster data. All parameters will be divided only into two main groups: quantitative (measurable) and qualitative parameters. Raster data do not primarily distinguish particular features and they store geometric properties only implicitly. Topological relationships are not explicitly stored in data; they cannot be used for analyses. These facts will be taken into consideration.

3.1 Proposed Quality Characteristics and Parameters

The proposed set of characteristics and parameters for raster data quality evaluation follows. It respects existing standards but it takes into account more aspects and it focuses namely on parameters suitable for raster data. Thus, some characteristics, parameters and their description follows from [2], [3], ISO standards and EuroRoadS project [16], some of them were proposed by the authors previously ([17], [18]) and some of them are newly proposed by the authors.

**Quantitative characteristics and parameters:**

1. **Accuracy:**
   - Geometric accuracy – focused on absolute and relative position accuracy. It expresses the difference between the data and true locations, it should include both vertical and horizontal accuracy when required
   - Thematic accuracy – measures the difference of attribute values from the true values; e.g. it expresses the accuracy of raster classification
   - Temporal accuracy – accuracy of time measurement, validity in a time and time consistency are measured

2. **Level of detail**
   - Spatial resolution – describes size of the cells in relation to the reality (i.e. how large area is covered by the cell); it consequently influences size of the smallest objects that can be resolved in the imagery and raster data
   - Scale

3. **Completeness:**
   - Completeness – omission of features, their attributes or relationships is measured
   - Completeness – redundancy and excesses in data (data contain not necessary features and attributes)
   - Spectral resolution – number of spectral bands available; valid for remote sensing data only

4. **Correctness** – expresses the extent of conformity of data with the reality; a low level of correctness can be for example caused by low accuracy or not up-to-dateness. Parameters:
   - Geometric correctness
   - Thematic correctness
   - Temporal correctness – temporal resolution in the case of remote sensing data

5. **Consistency** – level of observing logical (conceptual) schemas and models of reality. Topological consistency is not applicable for raster data. Parameters:
   - Geometric consistency
   - Thematic consistency

6. **Up-to-dateness** – describes how quickly data follows changes of the real world in time. Parameters:
   - Frequency of updates
   - Date of the last updates
7. **Availability** – valid mostly for online data or shared data which are stored on the server side:
   - Failure rate – frequency with which data are not accessible (measured in percents or in seconds, minutes or hours per given period)

**Qualitative characteristics and parameters:**
1. **Source:**
   - Source of data – data can come from trustworthy sources (e.g. national statistical office, public administration authority) or from not well-known source, e.g. from a Web site without clear responsibility for data
2. **Accessibility**
   - Data format – the best data format is the format which is regularly used
   - License – describes author rights and conditions under which data are available
   - Costs – high costs can limit some users
3. **Aim**
   - Purpose of data creation – intention to their future use
4. **Utilization**
   - Real utilization of data – references to solving which problems data were used
5. **Coordinate system**
   - Three various coordinate systems are used in the Czech Republic. The cartographic distortion can cause problems in precision.

### 3.2 Utilization of Proposed Quality Parameters – Digital Elevation Model

Digital elevation model (DEM) will be used in this example to evaluate aspect, slope and calculate hillshade to evaluate amount of sunlight.

Task description: it is required to find suitable places for new vineyards in the given region. South-west or south-facing slope, at 250 – 300 m above sea level is expected. The slope itself should be about 10 - 25° to assure enough sunlight on one side and a possibility of cultivation on the other side [17]. The task will be solved just to show differences in obtained results when data of a good and poor level quality are used.

<table>
<thead>
<tr>
<th>Quality Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometric accuracy</td>
<td>min. 2 m</td>
</tr>
</tbody>
</table>

In this case only evaluation by stating the values was used. The next possibility is to use multi-criteria evaluation methods, e.g. reference point method [18], AHP, etc.

Data sets used for the example task solution:
- **ZABAGED** – aim: a digital equivalent of the Base map of the Czech Republic 1 : 10 000, source: Czech Office for Surveying, Mapping and Cadastre, example data set, year 2007, data format: SHP, contours: 5 m, license: free example
- **ArcCR500** – aim: data for students, a data set in scale 1 : 500 000, source: ARCDATA PRAHA, s.r.o., year 2002, data format: SHP, contours: 50 m, license: free for study purposes

The whole area of interest is shown in Fig. 4 and Fig. 5. The pictures show DEM and contours based on each of the data sets, i.e. input data are visualized. A corresponding hillshade is used as a background layer to improve visualization. The difference between the data sets can be seen easily. It is caused namely by different scales (spatial resolution). ZABAGED (Fig. 4) is much more detailed dataset, its horizontal accuracy is 1 – 10 m, and its vertical accuracy is 3 m at maximum. Spatial resolution of DEM was 2 m. ArcCR500 (Fig. 5) has...
accuracy 100 – 250 m and it provided DEM with spatial resolution 5 m [17]. Because of the study purpose, older example data sets were used which were available for free. Both data sets were complete, only topographical data were required. In the case of thematic mapping, completeness must be more deeply evaluated [18]. Coordinate system was the same in both cases: S-JTSK.

The first step in the example task was to calculate slope, aspect and solar radiation. Then only areas which meet or given condition were selected by means of map algebra. The last optional step was to verify results using ortophotomap. These steps are not described in the paper because of their length and low importance. Both data sets were treated by the same way, using the same tools and keeping their original spatial resolution during the whole analysis.

Obtained results, i.e. places suitable for vineyard and meeting the given conditions were identified using both data sets but the results were very different. Results are summarized in the Table 2 and visualized in the Fig. 6.

Table 2 – Comparison of the Results of Analyses [17]

<table>
<thead>
<tr>
<th>Data set</th>
<th>ZABAGED</th>
<th>ArcCR500</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of places</td>
<td>1 788</td>
<td>82</td>
</tr>
<tr>
<td>Total area</td>
<td>283 744 m²</td>
<td>232 044 m²</td>
</tr>
<tr>
<td>The smallest place</td>
<td>2.6 m²</td>
<td>16 m²</td>
</tr>
<tr>
<td>The largest place</td>
<td>28 625 m²</td>
<td>102 803 m²</td>
</tr>
<tr>
<td>Average size of place</td>
<td>158,7 m²</td>
<td>2 830 m²</td>
</tr>
</tbody>
</table>

Fig. 4 - Area of interest, input data: ZABAGED

Fig. 5 - Area of interest, input data: ArcCR500

Fig. 6 – Differences in results (authors, based on [17])

4 Conclusion

Decision-making can be significantly influenced by information systems as they should provide relevant information to support decision-making. Data of an adequate quality level must be available in information systems, otherwise outputs will not be correct. The need for high quality data resulted into many standards and data quality models.

General purpose data quality models cannot be directly used to evaluate spatial data quality because the models do not take into consideration a specific nature of spatial data. Thus, attempts to propose quality models for spatial data were made, e.g. like the one within EuroRoadS project [16].

The EuroRoadS project identified a major problem of ISO standards for spatial data quality evaluation – they do not propose just one quality characteristics for one quality phenomena [16] and...
they are focused directly on technical parameters of data. EuroRoadS project was much more detailed, it proposed clear set of characteristics but it was mostly focused on vector data, it omitted raster data.

This paper bridges the above mentioned gap – it proposed, within ISO standards framework, quality characteristics and parameters for raster spatial data. Namely, spatial resolution, spectral resolution, temporal resolution, source, price, license, data format were proposed, purpose of data creation.

In the end the paper demonstrated utilization of the proposed quality parameters – a brief description of an example case study was provided, including identification of appropriate values for the quality parameters.

Acknowledgments. The authors would like to thank the Grant Agency of the Czech Republic for funding this research project under the grants No. 402/09/0219 with title ‘Usability of software tools for support of decision-making during solving spatially oriented problems’ and 205/09/P120 with title ‘Use of geoinformation technologies for detection of places in town with high risk for handicapped people’.

References: