DOI: http://doi.org/10.22364/hssl.26.2.4

DATA QUALITY SCALE FOR DATA QUALITY ASSESSMENT: METHODOLOGICAL GUIDELINES AND PRACTICAL IMPLEMENTATION

Daina Šķiltere
Dr. oec.

Svetlana Jesiļevska
Dr. oec.

Abstract

In this article, the complex methodology for the entire data quality treatment – the *Data Quality Scale (DQS)* is proposed. The *Data Quality Scale (DQS)* is developed by Dr. oec. Svetlana Jesiļevska and Dr. oec. Daina Šķiltere. The *Data Quality Scale (DQS)* gives an opportunity to identify certain shortcomings of the quality of statistical data and to develop proposals to improve the quality of statistical data. The methodology consists of data quality dimensions and its definitions, indicators for assessment of data quality dimensions and experts' evaluations. The Data Quality Scale has good expansibility and adaptability as makes it possible to evaluate the quality of data at various levels of detail: at indicators' level, at the level of dimensions, to determine the entire quality of data. The research results enrich the theoretical scope of statistical data quality and lay a solid foundation for the future by establishing an assessment approach and studying evaluation algorithms.

Keywords: data quality, data quality dimensions, Data Quality Scale

Introduction

The authors of the Data Quality Scale have a wide experience in data quality assessment and have a profound background regarding data quality. Since 2012, the authors have been dealing with data quality issues. Data quality assessment and improvement are topical issues nowadays; the authors identified a number of problems of statistical data, such as data sources, recent data, regularity, timeliness of data, updating data, time series, data frequency, data comparability, data costs etc. Sometimes, the required data does not exist; data from different sources are not always comparable (Šķiltere and Jesiļevska, 2014). It is therefore of vital importance that a systematic approach is available to assess the quality of statistical data. The problem here is associated with selecting appropriate criteria to evaluate the goodness of statistical data, therefore, not just

related to the research paradigm and intention, but also to the beliefs held by both researchers and research participants (Šķiltere and Jesiļevska, 2014). Based on existing theory, the authors developed a system of quality indicators to be used to determine the quality of statistical data. This systematic approach consists of the following quality characteristics: data completeness, representativity, objectivity, quality of methodology, coherence, accessibility, accuracy of estimates, actuality, interpretability, statistical disclosure control, and optimal use of resources, utility, and informativeness. To some of the proposed data quality dimensions not much attention has been paid previously. The authors conducted an expert's survey to find out the most essential data quality dimensions. The set of data quality dimensions has been tested with experts using four different data usage contexts: data for scientific research, data for decision-making, data for analysis of the progress of a research object during the reporting period, and data for research object modelling and forecasting (Jesiļevska, 2017).

The authors found out that one of the most problematic data quality dimensions is data accuracy. In the scientific literature, many methods have been proposed to identify outliers for empirical distributions, such as Dixon Test, Grubbs Tests, Hampel's Test, Quartile Method, Nalimov Test, Walsh's Test, and Discordance Outlier Test etc. In 2010, Šķiltere D. and Danusēvičs M. developed a method to assess total errors of the truly non–linear trend models (Šķiltere and Danusēvičs, 2010). However, no method was available in the scientific literature for identifying outliers by analysing changes in the indicator under the influence of one or several factors. In 2015, Jesiļevska S. developed the Iterative method for reducing the impact of outlying data points. The Iterative method was awarded the 3rd Prize in the 2015 International Competition the IAOS Prize for Young Statisticians and was published in the Statistical Journal of the IAOS: Journal of the International Association for Official Statistics in 2016 (Jesiļevska, 2016).

Based on the previously developed integrated approach to data quality assessment that consists of 13 data quality dimensions and the assessment indicators for each dimensions, in this article the authors present a complex methodology for the entire data quality treatment – the Data Quality Scale (DQS).

Data quality, data quality dimensions and assessment indicators

Since all types of research must respond to the agreed canons of quality (Marshall and Rossman 2006), one cannot avoid discussing them, despite their philosophical and practical complexity, as well as the difficulty in

defining what quality means or covers. Nowadays there are multiple separate ways to define data quality and there is currently no commonly agreed definition on what data quality is. Different analysts and different agencies provide different answers (Brackstone 1999, Carson 2000, Pipino et al. 2002), but all agree that the quality evaluation is by its nature a multidimensional problem (Madnick S. et al, 2009; Wang R. and Strong D., 1996).

The data quality dimension is a characteristic for classifying data quality requirements. In fact, data quality dimensions provide an opportunity for assessment and control of data quality (Wang R. Y., Ziad M., Lee Y. W., 2001). The data quality literature provides a thorough classification of data quality dimensions; however, there are several discrepancies in the definition of most dimensions due to the contextual nature of quality. The six most important classifications of quality dimensions are provided by several scientists (Wand and Wang, 1996; Wang and Strong, 1996; Redman, 1996; Jarke et al., 1995; Bovee et al., 2001; Naumann, 2002). By analysing these classifications, it is possible to define a basic set of data quality dimensions, including accuracy, completeness, consistency, and timeliness, which constitute the focus of most authors (Catarci and Scannapieco, 2002).

Data quality is multi-dimensional; however, the most frequently mentioned dimensions are accuracy, completeness, consistency and timeliness. The choice of data quality dimensions is based on knowledge, intuitive understanding (Ballou D. P. and Pazer H. L., 1985), experience (Firth C. P. and Wang R. Y., 1996), and findings from scientific literature (Kriebel C. H., 1979). However, scientific research results (Wang R. Y., Storey V. C., Firth C. P., 1995) show that there is no agreement on data quality dimensions.

The scientific literature highlights the significance of systematic, science-based data quality assessment. Scientists indicate that it is important to assess quality within three levels: quality of data collection process, quality of final data and quality of data use. The following data quality dimensions mainly qualitative assessments are provided by researchers: data actuality, optimal use of resources, data interpretability, data coherence, data objectivity, quality of the data collection and processing methodology, data availability, informativeness, and utility. In addition, the methodology for assessment of these data quality dimensions is not fully developed:

- **Data representativeness** – Mainly, quantitative assessment methods, e.g. design effect, effective sample size-neff etc. Indicator – respondence level etc.
- **Data accuracy** – Mainly, quantitative methods for identification of outliers for empirical distribution, e.g., extreme value test, discordance test, Grubb's test, Dixon test etc.

- **Data completeness** – Simple factor method
- **Data actuality** – Data lag, information float, volatility
- **Data coherence** – Mainly, qualitative assessment approach
- **Data interpretability** – Mainly, qualitative assessment approach, e.g., survey of data users
- **Data utility** – Mainly, qualitative assessment approach, e.g., survey of data users
- **Quality of data collection and processing methods** – Mainly, evaluate in the perspective of data quality dimensions
- **Data availability** – Mainly, qualitative assessment approach, e.g., survey of data users
- **Data objectivity** – Not found

Two-tier system of indicators on data quality assessment

The authors developed a two-tier system of indicators for data quality assessment, which includes 13 data quality dimensions:

Data objectivity – *the ability of the initial data* to reflect the actual situation*

1. The compliance of the implementation of the specially organised statistical observation with the scientifically based methodology
2. The compliance of the implementation of the survey (as a method of statistical observation) with the scientifically based methodology
3. The adequacy of the number of questions asked to respondents to obtain the information necessary for data users
4. Providing initial data* stability in time (for example, the respondent's answers are based on opinions, judgments, ideas that are considered true)
5. Ensuring minimization of impact of numerous factors on the respondent's answers in the questionnaire:
   - impact of external events (e.g. political) on the initial data*
   - influence level of mentality (e.g., religion, culture, history, traditions) on the respondents' answers
   - the impact of public opinion on respondents' answers
6. Ensuring equal survey question understanding among statisticians and respondents, the question is asked unambiguously

Data completeness –*sufficiency of the initial data* to meet user needs*

1. Ensuring collection of all the initial data* that are needed to carry out the assessment of the phenomena:
   - in dynamics
   - by objects (industry, regions etc.)

Data representativity – *sample data generalization capabilities*

1. Ensuring sample planning according to the tasks of the statistical research
2. Ensuring sampling planning component – sample size according to the tasks of the statistical research
3. Sufficiency of the survey response rate to fulfil tasks of statistical research
4. Ensuring the minimum number of incorrect answers (e.g. incomplete, illogical, not corresponding to reality) obtained within the survey

Data accuracy – *data meets the factual situation (data are free of error, correct)*

1. Implementation of systematic evaluation and correction for
   - initial data* and interim results
   - mistakes that may occur during the data collection and processing process (sampling errors and non-sample errors)
2. Evaluation of methodology for calculating derivative statistical indicators*****
3. Performing data audits in accordance with internationally recognized and scientifically valid procedures and data audit guidelines
4. Performing data correction in the case of changes in the subject of the study (data correction, recalculation)
5. Clarification of preliminary statistical indicators**** in accordance with well-tested and clearly understandable procedures

Quality of methodology – *scientific justification of methodology (including approbation of methodology), correct use of methodology and unification level of methodology*

1. Regularity in performing:
   - evaluation of the quality of statistical studies
   - supervision and improvement of scientifically sound data collection and processing methodology
   - monitoring and improving the scientifically-based methodology for calculating derivative statistical indicators*****
2. Compliance of the data collection and processing methodology with EU and international criteria
3. Evaluation of the results of the testing of the survey questionnaire before the statistical survey
4. Coherence of the data collected during the data collection and processing with the needs of the main data users (mainly, government institutions)
5. The relevance of data collection and processing processes to the rapidly changing environment

6. Opportunity for the operative implementation of new methods of data collection and processing and / or introduction of methodologies on new indicator calculation
7. Ensuring the level of unification of the data collection and processing methodology
8. Balancing the amount of resources invested in complex indicators (such as the European Innovation Scoreboard) with the utility of these complex indicators

Data coherence – *logical links between different statistical surveys' results, the data from various sources are comparable*

1. Coherence of methodology (definitions, classifications, methods) between statistical domains (different economic and social spheres, etc.)
2. Use of micro-data from one survey to improve the quality of another survey
3. Compliance of trends of correlating indicators within different statistical surveys
4. Collaboration with database maintainers to ensure data quality

Data actuality – *speed and frequency of renewal of data collection and processing*

1. Systematicity of monitoring of statistical data topicality and practical utility
2. Timeliness of the timing and the publication of statistical data, considering of the timing of publication of statistical indicators****
3. The relevance and adequacy of statistical data to needs of data users
4. Systematicity of statistical data renewal
5. The possibilities for reducing the period between the end of the reporting period and the publication of provisional*** / final data
6. Reduction of the period in the dynamics between the end of the reporting period and the publication of provisional*** / final data in comparison with the previous statistical surveys

Data accessibility – *simplicity of data availability to the users*

1. Providing access to statistical data for various categories of data users, respecting confidentiality requirements
2. The application of strict confidentiality requirements to external data users who have access to microdata** for research purposes
3. Implementation of a multitude ways of data dissemination: printed, files, CD-ROMs, Internet databases, etc.
4. Quality indicators are available for data users according to the European Statistics quality criteria

Data interpretability – *statistical data collection and processing methodology is available to the data users to make the correct interpretation of data*

1. Providing access to data collection and processing methodology for data users
2. Providing access for data users to definitions, calculation methodology, classifications etc. on socio-economic etc. indicators
3. Providing access for data users to interpretation of dynamic statistical indicators***** (e.g., growth rate, etc.)
4. Schematic presentation of components of complex indicators that enables data users to understand the nature of the indicator

Data informativeness –*data presentation form that enables data users to capture data quickly and easily navigate the data range*

1. Providing the possibility for data users to make an analysis of the data
2. Providing the possibility to create data tables online using interactive databases
3. Providing the possibility for data representation in interactive maps (selecting different territorial cuts of the country, only a specific part of the national territory, displaying data in comparison with other countries, etc.)
4. Ensuring possibilities of using interactive databases for creating tables online

Data utility – *data users' demand to the data*

1. Ensuring the possibility to use data:
   - for different purposes (for decision-making, research, forecasts, etc.)
   - by different users' categories (government, researchers, organizations, media etc.)
2. Systematicity of conducting analysis of data users' demand for data
3. Level of data users' satisfaction

Statistical disclosure control – *confidentiality of the information provided by respondents*

1. Ensuring of statistical confidentiality is stated in the law
2. Ensuring availability of the confidentiality policy to the public
3. Implementation of physical, technical and organizational measures in the statistical office to ensure the security of statistical databases

Optimal use of resources – *efficient use of existing resources for data collection and processing*

1. Ensuring a maximum use of potential of productivity of information and communication technologies during data collection, processing and dissemination

2. Performing various measures to improve the potential of administrative data for statistical purposes and to avoid direct surveys
3. Implementation of standardized solutions that improve the efficiency and productivity of resources used
4. Analysis and control of the amount of resources used (time, work, finances, etc.) in the process of collecting and processing data

* ***Initial data*** *– original raw data collected during the statistical study (statistical observation or survey).*

** ***Microdata*** *– tested and specified initial data used for the calculation of the preliminary and final data.*

*** ***Provisional data values*** *– data that require further clarification.*

**** ***Statistical indicators*** *– quantitative characteristics of changes of the phenomenon or object.*

***** ***Derivative statistical indicators*** *– are the quantitative and qualitative characteristics of a phenomenon or group of objects (for example, absolute aggregates, growth rate, average, proportion, etc.) based on a scientifically valid calculation method.*

The data quality dimensions proposed by the authors are essential during every stage of producing statistical data, ensuring systemic approach towards data quality assessment:

1. **Stage. Evaluation of the need for data** – Optimal use of resources
2. **Stage. Statistical data production process planning and development** – Quality of methodology, coherence of methodology, optimal use of resources
3. **Stage. Data collection** – Quality of methodology, coherence of data and methodology, accuracy, representativity, objectivity, actuality, statistical disclosure control, optimal use of resources
4. **Stage. Data processing** – Quality of methodology, coherence of data and methodology, accuracy, representativity, actuality, statistical disclosure control, optimal use of resources
5. **Stage. Data analysis** – Quality of methodology, coherence of data and methodology, accuracy, actuality, optimal use of resources
6. **Stage. Data dissemination** – Accessibility, informativeness, interpretability, utility, completeness, actuality, statistical disclosure control, optimal use of resources
7. **Stage. Data archiving** – Quality of methodology, coherence of data and methodology, statistical disclosure control, optimal use of resources, and
8. **Stage. Statistical data collection process evaluation** – Optimal use of resources

During the research, in cooperation with the Central statistical bureau of Latvia an expert survey of highly qualified specialists responsible for collection, processing and analysis of statistical information was carried out. In the experts' survey participated 19 experts from National statistical offices

representing the following countries: *Belgium, Armenia, Cyprus, Finland, Iceland, Czech Republic, Malta, Bulgaria, Romania, Slovak Republic, Ukraine, Lithuania, Belarus, Azerbaijan* and *Latvia*. The authors asked the experts to estimate the optimal level of significance of each data quality indicator in % corresponding to the theoretical guidelines of statistical science according to the following scale 0%→70%, 70%→90%, 90%→100%, 100%. For the multidimensional case, the authors propose to evaluate independently each indicator for assessment of data quality dimensions. Based on indicators' from the expert assessments the authors calculated the Dimension mean:

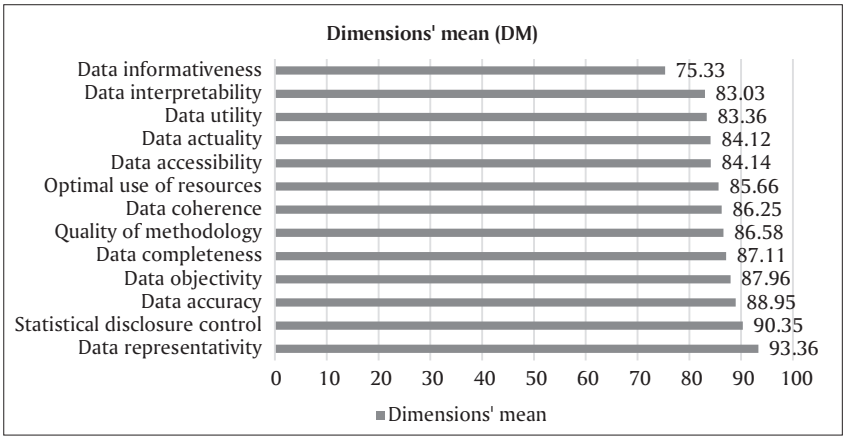

Figure 1. Data quality dimensions' mean (DM) and range according to the experts' evaluations

Source: prepared by the authors

In the experts'-statisticians view five the most important data quality dimensions are data representativity, statistical disclosure control, data accuracy, data objectivity and data completeness (see Figure 1).

In 2016, the authors conducted an expert survey in which both statisticians and data users were involved. Interesting to mention that the most significant differences in the answers of statisticians and users were identified in the context of data used for analysis the progress of research object during the reporting period. Statisticians consider quality of methodology as the most important dimension; data users argue that the most essential dimension is completeness. Data users put quality of methodology on the bottom of the list and believe that quality of methodology, accessibility and interpretability are equally less important dimensions (Jesiļevska S., 2017).

Data Quality Scale

The proposed method Data Quality Scale makes it possible to evaluate the quality of data at various levels of detail: at indicators, at the level of dimensions, and to determine the overall quality of data.

One key challenge is to determine what level of data quality is acceptable (or "good enough"). Based on indicators' expert assessments we calculated the Dimension mean and determined limit values for low, average and high-quality data (see Table 1 and Figure 2).

Table 1. Data Quality Scale. Limit values for data quality treatment according to the experts' evaluations

Dimensions	Limit values		
	For low quality data	For average quality data	For high quality data
Data objectivity	less than 67%	67% – 84%	84% – 100%
Data completeness	less than 64%	64% – 87%	87% – 100%
Data representativity	less than 77%	77% – 90%	90% – 100%
Data accuracy	less than 68%	68% – 87%	87% – 100%
Quality of methodology	less than 70%	70% – 80%	80% – 100%
Data coherence	less than 63%	63% – 84%	84% – 100%
Data actuality	less than 67%	67% – 79%	79% – 100%
Data accessibility	less than 54%	54% – 80%	80% – 100%
Data interpretability	less than 51%	51% – 79%	79% – 100%
Data informativeness	less than 58%	58% – 68%	68% – 100%
Data utility	less than 57%	57% – 77%	77% – 100%
Statistical disclosure control	less than 59%	59% – 84%	84% – 100%
Optimal use of resources	less than 66%	66% – 78%	78% – 100%
Total Data Quality Value	**81%** → 100%		

Source: prepared by the authors

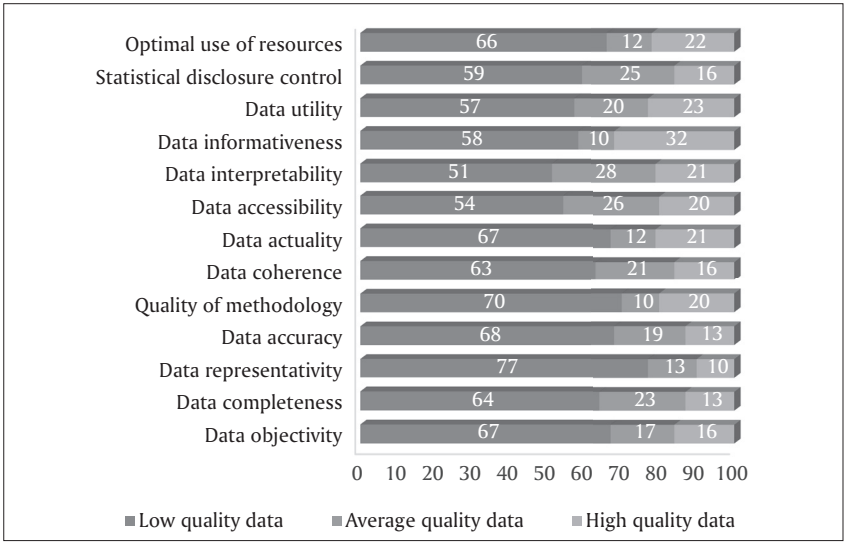

Figure 2. Data Quality Scale. Limit values for data quality treatment

Source: prepared by the authors

Low quality data is a problem for decision-making both in the country and companies' level, statistical data of low quality represent a significant cost factor for many companies, which is supported by findings from several surveys from industrial experts (Marsh, 2005). Kim and Choi (2003) who state, "There have been limited efforts to systematically understand the effects of low quality data. The efforts have been directed to investigating the effects of data errors on computer-based models such as neural networks, linear regression models, rule-based systems, etc." and "In practice, low quality data can bring monetary damages to an organization in a variety of ways". According to Kim (2002), the types of damage that low quality data can cause depend on the nature of data, the purpose of the use of data, the types of responses to the damages, etc. As a result, it is significant to identify data quality dimensions of low quality and to develop the ways in which these weaknesses could be improved.

The following main phases characterise the methodology:

1. data quality assessment on the level of the data quality indicators for a certain statistical data according to the following scale 0%→70%, 70%→90%, 90% 100%, 100%;
2. calculation of the Dimension mean and evaluation of the entire data quality level;
3. comparison with the optimal data quality level values (see Table 1),

4. identification of the shortcomings during the process of data collection on the level of data quality indicators,
5. validation and processing based on the assessment of the data quality indicators, and
6. choice of the optimal data quality improvement process.

Conclusions

The Data Quality Scale and the methodology can be used by the statisticians to understand the statistical data quality assessment and the various quality exchanges inside it. The authors are convinced that the Data Quality Scale will help statisticians to determine shortcomings of the data, to improve data quality significantly to improve the process of decision-making based on statistical data.

Having at his or her disposal a methodology of evaluating not only at the data quality dimensions' level, but also the entire statistical data quality, makes possible to use the Data Quality Scale for data form different areas of industry, to make data assessment on dynamics with the purpose to realise the progress in data quality, and to find out systematic failures of data collection, processing, validation etc.

To solve data quality problems effectively, both data users and data producers must use sufficient knowledge about solving data quality problems appropriate for their process areas. At minimum, statisticians must know what kind of data, how (this question includes mainly methodological issues), and why to collect the data; data users must know what data, how (what kind of analysis), and why (intended purpose) to use the data. In sum, the two main actors mentioned above have roles in a data production process and should cooperate closely to improve statistical data quality. Involvement of both statisticians and data users in the process of identifying and solving possible drawbacks of data opens new avenues for future research and practice.

REFERENCES

1. Ballou, D. P., Pazer, H.L. (1985), Modeling data and process quality in multi-input, multioutput information systems. *Management Science 31(2),* 150–162.
2. Bovee, M., Srivastava, R., Mak, B. A. (2001), Conceptual framework and belief-function approach to assessing overall information quality. In *Proceedings of the 6th International Conference on Information Quality*.
3. Brackstone, G. (1999), Managing Data Quality in a Statistical Agency. *Survey Methodology 25*, 139–149.
4. Carson, C. S. (2000), *What is Data Quality? A Distillation of Experience. Statistics Department.* International Monetary Fund.

5. Catarci, T., Scannapieco, M. (2002), Data quality under the computer science perspective. *Archivi Computer* 2.
6. Firth, C. P., Wang, R. Y. (1996), *Data Quality Systems: Evaluation and Implementation*. London, Cambridge Market Intelligence.
7. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P. (1995), *Fundamentals of Data Warehouses.* Springer Verlag.
8. Jesiļevska, S. (2016), Iterative method for reducing the impact of outlying data points: Ensuring data quality. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics 32(2),* 257–263.
9. Jesiļevska, S. (2017), Data quality dimensions to ensure optimal data quality. *The Romanian Economic Journal XX(63),* 89–103.
10. Kim, W., Choi, B. (2003), Towards Quantifying Data Quality Costs. *Journal of Object Technology 2(4),* 69–76.
11. Kim, W. (2002), On Three Major Holes in Data Warehousing Today. *Journal of Object Technology 1(4),* 39–47.
12. Kriebel, C. H. (1979), *Evaluating the quality of information systems*. Design and Implementation of Computer Based Information Systems. N. Szysperski and E. Grochla. Ed. Sijthtoff & Noordhoff, Germantown.
13. Madnick, S., Wang, R., Lee, Y., Zhu, H. (2009), Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality 1(1).*
14. Marshall, C., Rossman, G. B. (2006), *Designing Qualitative Research (4 ed.).* Thousand Oaks, CA: Sage.
15. Marsh, R. (2005), Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management 12(2),* 105–112.
16. Naumann, F. (2002), *Quality-driven query answering for integrated information systems*. Lecture Notes in Computer Science 2261.
17. Pipino, L. L., Lee, Y. W., Wang R.Y. (2002), Data Quality Assessment. *Communications of the ACM 45,* 211–218.
18. Redman, T. (1996), *Data Quality for the Information Age*. Artech House.
19. Šķiltere, D., Danusēvičs, M. (2010), Interval Forecasting Methods In Longterm Statistical Forecasting, *A Journal of the International Institute for General Systems Studies 11(1),* 11–20.
20. Šķiltere, D., Jesiļevska, S. (2014), Data quality evaluation in statistical data processing, *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics 30(4),* 425–430.
21. Šķiltere, D., Jesiļevska, S. (2014), Examining Dimensions of Data Validity, *International Journal of Statistics and Economics 15(3),* 18–24.
22. Wang, R. Y., Ziad, M., Lee, Y. W. (2001), *Data Quality*. New York: Springer.
23. Wang, R. Y., Storey, V. C., Firth, C. P. (1995), A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering 7(4*), 623−640.
24. Wang, R., Strong, D. (1996), Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems 12(4).*
25. Wand, Y., Wang, R. (1996), Anchoring data quality dimensions in ontological foundations. *Comm. ACM 39(11).*