# Iterative data quality management system

Svetlana JESIĻEVSKA[1]; Daina ŠĶILTERE[2]

***Abstract.*** *High-quality data are the precondition for analyzing and using statistics and for guaranteeing the value of the data. In this paper, the Iterative data quality management system is proposed. The methodology consists of two methods developed by the authors - the Iterative method for the reducing the impact of outlying data points in 2015 and the Data Quality Scale in 2018. The novelty of the Iterative method for the reducing the impact of outliers is the following: an iterative approach for determining the outlying data points is proposed; outliers are determined considering the impact of conjoined factors; estimation of weight coefficients of the outliers and estimation of the total measurement error of the non-linear regression model is carried out. The Iterative method received the Young Statistician Prize of the International Association for Official Statistics (IAOS) in 2015. 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, and to determine the entire quality of data. The Data Quality Scale gives an opportunity to identify certain shortcomings of the quality of statistical data and to develop proposals to improve the quality of the data. The research results enrich the theoretical scope of the 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, Iterative method for reducing the impact of outlying data points*

## 1. Introduction

Authors of the Data Quality Scale have a broad experience in data quality assessment and have a profound background on the data quality. Since 2012, authors have been dealing with data quality issues. Data quality assessment and improvement are topical issues nowadays; authors identified plenty of sources of quality problems of statistical data like data sources, regularity, timeliness of data, updating data, time series, data

[1] Dr.oec; *Eurointegration and economic development; E-mail: euroedpu@gmail.com*
[2] Dr.oec., *University of Latvia E-mail: daina.skiltere@lu.lv*

frequency, data costs etc. Sometimes, the required data do not exist; data from different sources are not always comparable (Šķiltere & Jesiļevska, 2014). It is therefore of vital importance that a complex approach is available to assess the quality of the statistical data. The problem here is associated with selecting appropriate criteria to evaluate the goodness of the 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 & Jesiļevska, 2014). Based on existing theory, authors developed a system of quality dimensions to determine the quality of statistical data. This systematic approach consists of the following data quality dimensions: data completeness, representativity, objectivity, quality of methodology, coherence, accessibility, accuracy of estimates, actuality, interpretability, statistical disclosure control, optimal use of resources, utility, informativeness. To some of the proposed data quality dimensions not much attention has been paid previously. 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 the progress of research object during the reporting period, data for research object modeling and forecasting (Jesiļevska, 2017).

Authors found out the 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, like Dixon Test, Grubbs Tests, Hampel's Test, Quartile Method, Nalimov Test, Walsh's Test, 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 & Danusēvičs, 2010). However, no method was available in the scientific literature for identifying outliers by analyzing changes in the indicator under the influence of one or several factors. In 2015 Jesiļevska S. developed Iterative method for reducing the impact of outlying data points. The Iterative method got 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, in this paper authors present the complex methodology for the entire data quality treatment – the Iterative data quality management system.

## 2. Data Quality Scale

The Data Quality Scale is based on the two-tier system of indicators on data quality assessment, which includes 13 data quality dimensions: data completeness, representativity, objectivity, quality of methodology, coherence, accessibility, accuracy of estimates, actuality, interpretability, statistical disclosure control, optimal use of

resources, utility, informativeness and indicators for assessment of data quality dimensions (see Table 1).

**Table 1. Definitions and assessment indicators for data quality dimensions**

| Data quality dimensions and definitions | Indicators for assessment of 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 organized statistical observation with the scientifically based methodology<br>2. The compliance of the implementation of the survey (as a method of statistical observation) with the scientifically based methodology<br>3. The adequacy of the number of questions asked to respondents to obtain the information necessary for data users<br>4. Providing initial data* stability in time (for example, the respondent's answers are based on opinions, judgments, ideas that are considered true)<br>5. Ensuring minimization of impact of numerous factors on the respondent's answers in the questionnaire:<br>- impact of external events (e.g. political) on the initial data*<br>- influence level of mentality (e.g., religion, culture, history, traditions) on the respondents' answers<br>- the impact of public opinion on respondents' answers<br>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*<br>that are needed to carry out the assessment of the phenomena:<br>- in dynamics<br>- by objects (industry, regions etc.) |
| Data representativity − *sample data generalization capabilities* | 1. Ensuring sample planning according to the tasks of the statistical research<br>2. Ensuring sampling planning component - sample size according to the tasks of the statistical research<br>3. Sufficiency of the survey response rate to fulfill tasks of statistical research<br>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<br>- initial data* and interim results<br>- mistakes that may occur during the data collection and processing process (sampling errors and non-sample errors)<br>2. Evaluation of methodology for calculating derivative statistical indicators*****<br>3. Performing data audits in accordance with internationally recognized and scientifically valid procedures and data audit guidelines<br>4. Performing data correction in the case of changes in the subject of the study (data correction, recalculation)<br>5. Clarification of preliminary statistical indicators**** in accordance with well-tested and clearly understandable procedures |
| Quality of methodology − *scientific justification* | 1. Regularity in performing:<br>- evaluation of the quality of statistical studies<br>- supervision and improvement of scientifically sound data collection and |

| Data quality dimensions and definitions | Indicators for assessment of data quality dimensions |
|---|---|
| *of methodology (including approbation of methodology), correct use of methodology and unification level of methodology* | processing methodology<br>- monitoring and improving the scientifically-based methodology for calculating derivative statistical indicators*****<br>2. Compliance of the data collection and processing methodology with EU and international criteria<br>3. Evaluation of the results of the testing of the survey questionnaire before the statistical survey<br>4. Coherence of the data collected during the data collection and processing with the needs of the main data users (mainly, government institutions)<br>5. The relevance of data collection and processing processes to the rapidly changing environment<br>6. Opportunity for the operative implementation of new methods of data collection and processing and / or introduction of methodologies on new indicator calculation<br>7. Ensuring the level of unification of the data collection and processing methodology<br>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.)<br>2. Use of micro-data from one survey to improve the quality of another survey<br>3. Compliance of trends of correlating indicators within different statistical surveys<br>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<br>2. Timeliness of the timing and the publication of statistical data, considering of the timing of publication of statistical indicators****<br>3. The relevance and adequacy of statistical data to needs of data users<br>4. Systematicity of statistical data renewal<br>5. The possibilities for reducing the period between the end of the reporting period and the publication of provisional*** / final data<br>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<br>2. The application of strict confidentiality requirements to external data users who have access to microdata** for research purposes<br>3. Implementation of a multitude ways of data dissemination: printed, files, CD-ROMs, Internet databases, etc.<br>4. Quality indicators are available for data users according to the European Statistics quality criteria |
| Data interpretability − *statistical data collection and* | 1. Providing access to data collection and processing methodology for data users<br>2. Providing access for data users to definitions, calculation methodology, |

| Data quality dimensions and definitions | Indicators for assessment of data quality dimensions |
|---|---|
| *processing methodology is available to the data users to make the correct interpretation of data* | classifications etc. on socio-economic etc. indicators<br>3. Providing access for data users to interpretation of dynamic statistical indicators**** (e.g., growth rate, etc.)<br>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<br>2. Providing the possibility to create data tables online using interactive databases<br>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.)<br>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:<br>- for different purposes (for decision making, research, forecasts, etc.)<br>- by different users' categories (government, researchers, organizations, media etc.)<br>2. Systematicity of conducting analysis of data users' demand for data<br>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<br>2. Ensuring availability of the confidentiality policy to the public<br>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<br>2. Performing various measures to improve the potential of administrative data for statistical purposes and to avoid direct surveys<br>3. Implementation of standardized solutions that improve the efficiency and productivity of resources used<br>4. Analysis and control of the amount of resources used (time, work, finances, etc.) in the process of collecting and processing data |

*Source: prepared by authors*

** 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.*

Data quality dimensions proposed by authors are essential during every stage of producing statistical data, ensuring systemic approach towards data quality assessment (see Table 2):

**Table 2. Essential data quality dimensions within data quality preparation stages**

| Statistical data preparation stages | Data quality dimensions |
|---|---|
| *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 |
| *8. stage. Statistical data collection process evaluation* | Optimal use of resources |

*Source: prepared by authors*

During the research, the 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*. We kindly asked 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%. We asked to evaluate independently each indicator for assessment of data quality dimensions. Based on indicators' expert assessments we calculated the Dimension mean (see Table 3).

**Table 3. Data quality dimensions' mean (DM) and rang according to the experts' evaluations**

| Dimensions | Dimension mean (DM) | Rang (R) |
|---|---|---|
| Data objectivity | 87.96 | 4 |
| Data completeness | 87.11 | 5 |
| Data representativity | 93.36 | 1 |
| Data accuracy | 88.95 | 3 |
| Quality of methodology | 86.58 | 6 |
| Data coherence | 86.25 | 7 |

| Dimensions | Dimension mean (DM) | Rang (R) |
|---|---|---|
| Data actuality | 84.12 | 10 |
| Data accessibility | 84.14 | 9 |
| Data interpretability | 83.03 | 12 |
| Data informativeness | 75.33 | 13 |
| Data utility | 83.36 | 11 |
| Statistical disclosure control | 90.35 | 2 |
| Optimal use of resources | 85.66 | 8 |

*Source: prepared by authors*

The *Data Quality Scale* makes it possible to evaluate the quality of data at various levels of detail: at indicators' level, at the level of dimensions, the entire 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 4).

**Table 4. 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 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 to improve these weaknesses.

The following main phases characterize the methodology:

1. data quality assessment on the level of the data quality indicators (see the list of indicators on Table 1) 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 5.1),
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,
6. choice of the optimal data quality improvement process.

## 3. The Iterative method for the reducing the impact of outlying data points

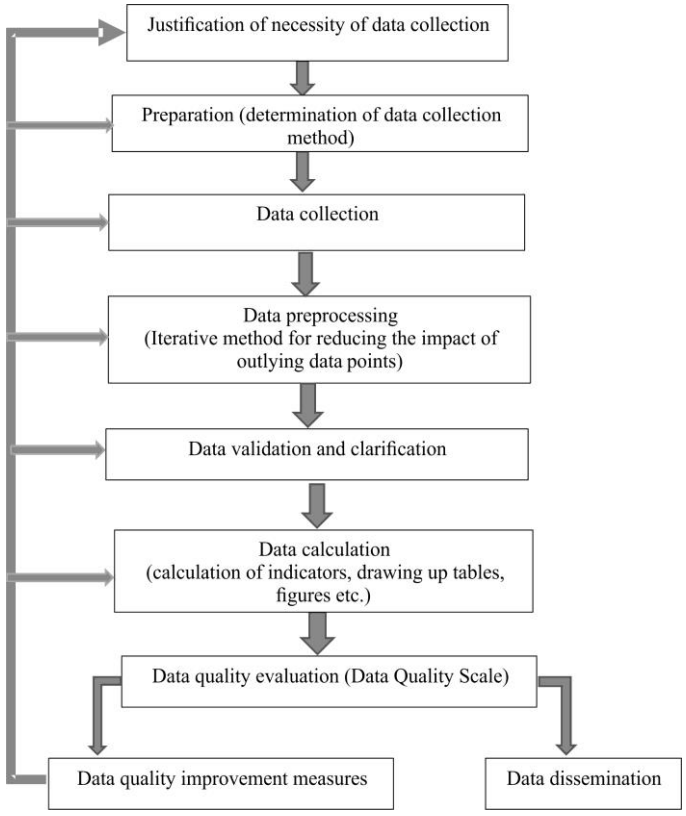
Detection and treatment of values deviating extremely from other values of the data set is an old problem of statistics. The aim of the outlier treatment is improving the estimation. The iterative method for the reducing the impact of outlying data points described aims to ensure statistical data quality during data pre-processing step. The Iterative method involves a series of steps to identify outlying data points and reduce its impacts. During the first step, an indicator and a factor is selected for further analysis. The conjoined factor may be used during the analysis. During the second step, the best-fit regression model for further analysis is determined. The third step is the total estimate error of the chosen regression model measurement. The fourth step is the most extensive as it consists of four substeps. During the fourth step, firstly, potential outlier points are identified. After potential outlier points are evaluated for the reason of their existence and factual outlier points are identified. Then, the author suggests to minimize the impact of factual outliers on the results: weight ratio for the outlier data point is determined by the normal distribution law. As a result, regression's model recalculation

with the corrected data is performed. The last step is validation. During the validation step, the received results should be analyzed and evaluated. If potential outliers are detected, come back to the 2nd step and run a new Iterative method circle.

## 4. The Iterative data quality management system

Based on the previous research results, developed the two-tier system of indicators on data quality assessment, the Iterative method for reducing the impact of outlying data points, the Data Quality Scale, authors propose the Iterative data quality management system (see Fig.1).

*Fig.1. Iterative data quality management system*




*Source: written by the authors*

## Conclusions

The Iterative data quality management system can be used by the statisticians to understand the statistical data quality assessment and the various quality exchanges inside it. We are convinced that the Iterative data quality management system 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.

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

Ballou DP, Pazer HL. Modeling data and process quality in multi-input, multioutput information systems. Management Science. 1985; 31(2): 150-162.

Bovee M, Srivastava R, Mak B. A conceptual framework and belief-function approach to assessing overall information quality. In Proceedings of the 6th International Conference on Information Quality. September 2001.

Brackstone G. Managing Data Quality in a Statistical Agency. Survey Methodology. 1999; 25: 139-149.

Carson CS. What is Data Quality? A Distillation of Experience.Statistics Department. International Monetary Fund. 2000.

Catarci T, Scannapieco M. Data quality under the computer science perspective. Archivi Computer 2; 2002.

Firth CP, Wang RY. Data Quality Systems: Evaluation and Implementation. London: Cambridge Market Intelligence; 1996.

Jarke M, Lenzerini M, Vassiliou Y, Vassiliadis P. Fundamentals of Data Warehouses. Springer Verlag; 1995.

Jesiļevska, S. "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, 2016, Vol. 32, No. 2, pp. 257-263

Jesiļevska, S. "Data quality dimensions to ensure optimal data quality", The Romanian Economic Journal, 2017, Year XX, No 63, pp. 89-103.

Kim W, Choi B. Towards Quantifying Data Quality Costs. Journal of Object Technology. 2003; 2(4): 69-76.

Kim W. On Three Major Holes in Data Warehousing Today. Journal of Object Technology. 2002; 1(4): 39-47

Kriebel CH. Evaluating the quality of information systems. Design and Implementation of Computer Based Information Systems. N. Szysperski and E. Grochla. Ed. Sijthtoff & Noordhoff, Germantown; 1979.

Madnick S, Wang R, Lee Y, Zhu H. Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality. 2009; 1(1).

Marshall C, Rossman GB. Designing Qualitative Research (4 ed.). Thousand Oaks, CA: Sage; 2006.

Marsh R. 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. 2005; 12(2): 105-112.

Naumann F. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science. 2002; 2261.

Pipino LL, Lee YW, Wang RY. Data Quality Assessment. Communications of the ACM 2002; 45: 211-218.

Redman T. Data Quality for the Information Age. Artech House; 1996.

Šķiltere, D., Danusēvičs, M., 2010, Interval Forecasting Methods In Longterm Statistical Forecasting, *A Journal of the International Institute for General Systems Studies*, Volume 11(1), pp. 11-20.

Šķiltere, D., Jesiļevska, S. "Data quality evaluation in statistical data processing", Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 2014, Vol. 30., No. 4., pp. 425-430

Šķiltere, D., Jesiļevska, S. "Examining Dimensions of Data Validity", International Journal of Statistics and Economics, 2014, Vol. 15, No. 3, pp.18-24

Wang RY, Ziad M, Lee YW. Data Quality. New York: Springer; 2001.

Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering. 1995; 7(4): 623−640.

Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 1996; 12(4).

Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. Comm. ACM. 1996; 39, 11.