Data Quality Metrics How to Measure and Improve Accuracy

Bharath Kishore Gudepu[1], Oscar Gellago[2], Rebecca Eichler[3]
[1]Senior Informatica Developer, Transamerica, 10100 N Central Expy Ste 595, Dallas, TX 75231
[2]University of Žilina, Žilina, Slovakia
[3]PRA Group Inc., USA

ABSTRACT

The quality of data, particularly its assessment, has been extensively examined in both research and practical applications. To facilitate economically driven management of data quality and decision-making under uncertainty, it is crucial to evaluate the data quality level using robust metrics. Nevertheless, if these measures are not well delineated, they may result in erroneous conclusions and financial detriment. Consequently, within a decision-oriented framework, we delineate five prerequisites for data quality measures. These requirements pertain to a measure designed to facilitate economically driven management of data quality and decision-making under uncertainty. We further illustrate the applicability and effectiveness of these requirements by assessing five data quality measures across several data quality dimensions. Furthermore, we examine the practical ramifications of implementing the outlined standards. The two most important requirements for data quality are consistency and accuracy. Database inconsistencies and errors are often caused by breaches of integrity requirements. To make a filthy database D consistent, automated techniques are required to locate a repair D0 that fulfils criteria and differs "minimally" from D. It's crucial to guarantee that the automatically generated fix D0 is accurate and makes sense. D0 should deviate from the "correct" data within a given range. This study explores practical approaches to improve data consistency and accuracy. We use conditional functional dependencies (CFDs) from [6] to ensure data consistency and detect mistakes that standard methods may miss. We present two techniques to increase data consistency: one for automatically computing a repair D0 that satisfies a set of CFDs, and another for progressively discovering a repair in response to clean database changes. We demonstrate that both challenges are insurmountable. We empirically validate the effectiveness and efficiency of our heuristic algorithms. We created a statistical strategy to ensure the accuracy of algorithmic fixes beyond a preset rate without requiring unnecessary human engagement.

Keywords: Data Quality Metrics, Data Accuracy, Data Quality, Data Governance, Data Management, Data Profiling, Metadata, Compliance, Data Cleansing, Analytics, Business Intelligence, Data Integrity, Enterprise Data, Data Improvement, Measurement

Introduction

The quality of data is essential for decision-making and gaining a competitive edge in the realm of big data. Organisations are more dependent on data to inform decision-making and secure a competitive edge. To make educated and successful judgements, it is crucial to evaluate and ensure the quality of the underlying data. The three attributes of big data—Volume, Velocity, and Variety—render the guarantee of data quality increasingly difficult. Substandard data quality incurs an annual cost of $3.1 trillion to the US economy.

Data quality metrics offer quantifications for data perspectives, with higher (lower) metric values indicating superior (worse) data quality, and each level of data quality denoted by a distinct metric value. They are essential for two primary reasons: firstly, metric values facilitate data-driven decision-making in uncertain conditions, and secondly, they underpin

an economically focused control of data quality. Data quality enhancement procedures should be implemented just when the advantages of improved data quality surpass the related expenses. To analyse the economic efficiency of data quality enhancement methods, robust data quality metrics are required to evaluate the changes in data quality levels [1-3].

To tackle this research question, five criteria for data quality metrics are proposed: the presence of minimum and maximum metric values (R1), the interval scaling of the metric values (R2), the integrity of the configuration parameters and the calculation of the metric values (R3), the robust aggregation of the metric values (R4), and the cost-effectiveness of the metric (R5). These standards are founded on a decision-centric paradigm that facilitates decision-making under uncertainty and promotes an economically driven approach to data quality management. Data quality measurements that fail to fulfil these criteria may result in erroneous choices and/or financial losses.

The necessity for these characteristics is also corroborated by discourse in other study domains, including software engineering. A comprehensive set of features for the rigorous definition of software measures, which researchers may utilise to test their novel metrics and which can be regarded as essential criteria for software metrics. The SQuaRE series under ISO/IEC standards is designed to aid developers and purchasers of software products in specifying and assessing quality criteria.

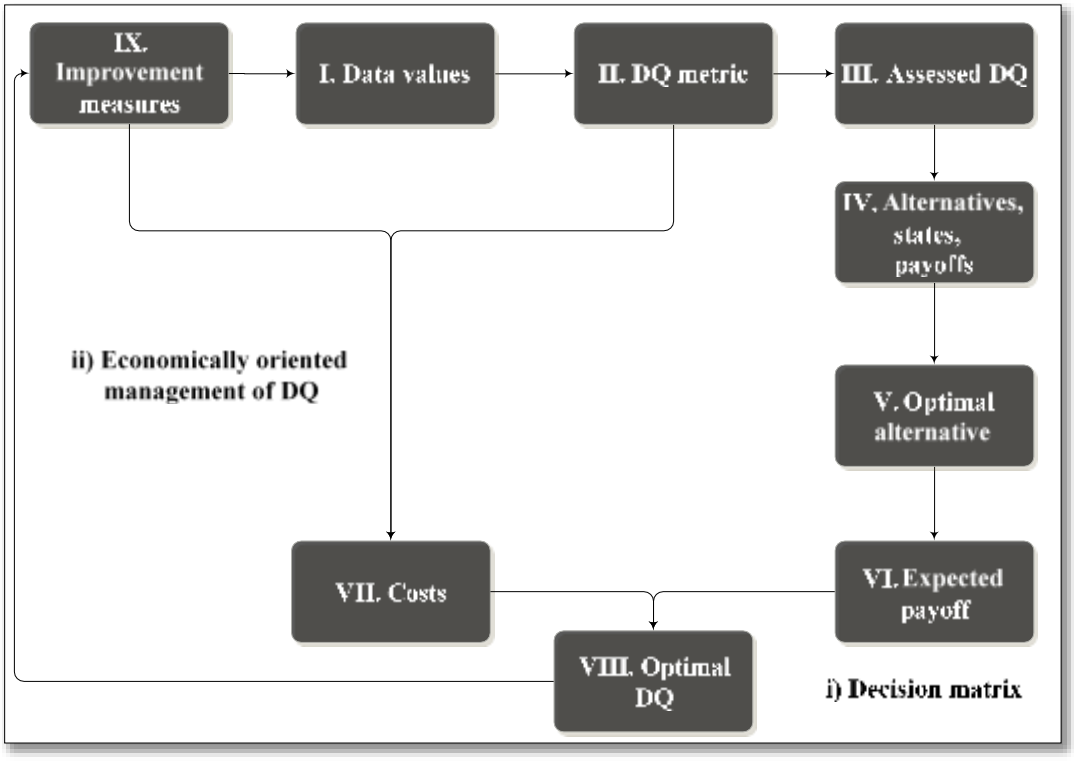

Figure1: Economically oriented management of DQ

Criteria For Data Quality Metrics

Requirement 1 (R1): Presence of Minimum and Maximum Metric Values

Group 1 asserts that data quality measures must assume values within a certain range. The majority of the criteria under this group, such as validity range and definition clarity, are ambiguously articulated, rendering them challenging to validate. Consequently, the significance of these conditions and the potential repercussions of their non-fulfillment remain ambiguous (e.g., measurability just asserts that the range must be discrete). To resolve these challenges, we suggest and substantiate the following requirement:

Requirement 1 (R1) (Presence of minimum and maximum metric values). The metric values must be constrained both above and below, capable of achieving a minimum (indicative of optimal bad data quality) and a maximum (indicative of optimal high data quality). Specifically, for any real-world value ωm, both minimum and maximum values must be achievable in relation to ωm [4-7].

Rationale. Initially, we will examine the subsequent statement (a), which will be referenced repeatedly throughout this justification:

There must be a singular metric value denoting optimal data quality and a singular metric value indicating suboptimal data quality. When a data quality metric is expressed as a mathematical function, (R1) stipulates that this function must be constrained both below and above, achieving a minimum and maximum value. Nevertheless, several established measures fail to reach a minimum or maximum, thereby resulting in erroneous assessments of choice alternatives (refer to III-VI in Figure 1). In such instances, it is not feasible to determine if the evaluated data quality level can or should be enhanced to facilitate improved decision-making (see VI-IX in Figure 1). Consequently, superfluous enhancement steps for data values of already exemplary quality may be undertaken, as the metric values fail to accurately reflect that optimal data quality has been achieved. Furthermore, when evaluating data quality repeatedly using a metric that fails to meet (R1), neither the comparability nor the validation (e.g., against a benchmark, such as a requisite completeness level of 90% of the analysed database) of the metric values across different evaluations is assured. Furthermore, when a particular data quality enhancement measure is implemented, there is no benchmark, in terms of minimum and maximum values, to assess the rankings over time (for instance, consider a user survey about the current data quality level without any guidance on the scale of values to be provided by the users). This opposes an economically driven approach to data quality management.

Requirement 3 (R3): Quality of Configuration Parameters and Determination of Metric Values

Group 4 includes standards indicating that a data quality metric must be adjustable to accurately represent the specific application environment. This, however, pertains just to one crucial component. Established scientific quality requirements, namely objectivity, reliability, and validity, must be met by data quality measures, although they have not been addressed in the existing literature. Furthermore, both the metric values and the configuration

settings of a data quality metric must adhere to these quality standards to prevent unsatisfactory outcomes (refer to II-III in Figure 1). To mitigate these issues, we suggest and substantiate the following requirement:

Requirement 3 (R3) pertains to the quality of setup settings and the assessment of metric values. It should be feasible to ascertain the configuration parameters of a data quality measure based on the quality criteria of objectivity, reliability, and validity. The same applies to the ascertainment of the metric values [8-11].

A substantial corpus of work addresses the quality requirements of objectivity, reliability, and validity in measurements. We will first succinctly examine these requirements in relation to data quality measures. Subsequently, we substantiate their significance using our decision-oriented approach.

The objectivity of the configuration settings and data quality metric values indicates the extent to which these parameters, values, and the methods for ascertaining them (e.g., SQL queries) are free from extraneous influences (e.g., interviewers). This criteria holds particular significance for data quality measures need expert evaluations to ascertain the configuration parameters or metric. Objectivity is compromised if estimations are derived from an insufficient number of experts or if external factors, such as the specific conduct of the interviewers, are not mitigated. Objectivity is compromised when metrics do not provide a clear statement of valid techniques for determining the relevant parameters and values. In this instance, measurements may have varying outcomes if used repeatedly. To prevent subjective outcomes and guarantee objectivity, the data quality metric and its configuration parameters must be clearly specified and established by objective techniques, such as statistical methods [12-17].

The reliability of measurement pertains to the precision with which a parameter is assessed. Reliability conceptualises the reproducibility of the outcomes derived from the methodologies employed to ascertain the configuration parameters or metric values. Methods will lack reliability if they use expert estimations that fluctuate over time or vary across various groups of experts. Reliability may be assessed by the correlation of findings derived from several metrics. Consequently, data quality measures that depend on expert assessments must establish a dependable methodology for ascertaining the configuration parameters and metric values. To assure the dependability of configuration settings and metric values, accurate database searches or statistical approaches may be employed. The outcome of the corresponding procedure remains consistent when performed repeatedly to the same data.

Validity is defined as the amount to which a metric accurately measures what it claims to measure or the degree to which it assesses the theoretical concept of interest. Thus, the validity of a method for ascertaining configuration parameters or metric values pertains to the extent of precision with which the method accurately measures its intended target. Typically, the validity of determining a configuration parameter or metric value is compromised if the determination contradicts its intended purpose. Numerous examples

demonstrate the practical significance of validity within the framework of data quality measures. The timeliness measure incorporates the configuration parameter $Currency$, which signifies "the promptness of data updates." The mathematical statement $Currency = Age + (DeliveryTime - InputTime)$ appears to contradict this objective assert that a metric value of zero signifies that "each validated data object contains at least one critical defect." Nevertheless, the mathematical definition of the metric indicates that for it to be zero, each data object must encompass all conceivable significant faults. Validity can be attained by consistent definitions, database queries, or statistical calculations designed to ascertain the relevant parameter or value in accordance with its specification. Moreover, limiting the application scope of a measure enhances its validity [18-20].

Application of The Requirements

We illustrate the applicability and effectiveness of our criteria by assessing five measures from the literature. We selected these metrics encompassing timeliness, completeness, reliability, correctness, and consistency to offer a comprehensive perspective on various dimensions of data quality and to demonstrate that the specified requirements can be effectively applied to multiple dimensions of data views and values stored within an information system. To enhance the transparency and comprehensibility of the metrics evaluation, we reference the following application context, a corporation must determine which current customers, such as corporate clients, to approach with a new product offer in a CRM mailing campaign based on recorded data. The two choice possibilities for the corporation about each consumer in the database are $a1$: to include the customer in the campaign or $a2$: to exclude them. The potential states of nature, contingent upon a specific likelihood of acceptance, are $s1$: the consumer accepts, or $s2$: the customer rejects the offer. The advantages of implementing a data quality metric in this situation are significant.

Practical Implications

This section addresses the significance and prioritisation of the needs, emphasising their practical consequences. We offer a comprehensive analysis for (R1) and (R2), along with distinct talks for (R3), (R4), and (R5). Table 8 encapsulates the findings.

R1: Presence of Minimum And Maximum Metric Values R2: Interval-Scaled Metric Values

(R1) and (R2) are especially pertinent when decisions on various data quality enhancement strategies, or more broadly, decision alternatives based on economic factors, are made according to the metric values (cf. economically orientated management of data quality). Specifically, let us assume that the objective of a certain application is to assess the currency of two data values of an attribute and to determine if the first data value is more current than the second (i.e., to formulate a true/false statement). In this particular instance, a straightforward ranking of the metric values for the currency of the two data points would be adequate. This analysis does not concern itself with the magnitude of the disparity between the currency metric values of the two data points, nor is it necessary to ascertain if

the interpretation of either or both currency metric values indicates (very) current or obsolete data values.

Nonetheless, for the vast majority of real applications, such a simplistic evaluation as a true/false statement is inadequate. A judgement on several alternatives, evaluated by economic criteria, must be made based on the metric values. In such instances, if just a ranking is accessible, validation against a designated benchmark (e.g., a requisite completeness level of 90% of the evaluated database) becomes unfeasible, hindering the metric's applicability for decision-making. Moreover, a ranking cannot substantiate the choice for the enhancement of the assessed data quality level based on economic factors, nor can it determine the feasibility of such an increase. Furthermore, employing such a metric renders the comparison between the outcomes of a data quality enhancement method and its associated costs ambiguous. All of these factors are essential for an economically focused control of data quality.

In summary, a measure may be particularly formulated for the purpose of evaluating the ranks of current data quality levels or utilised just within that context. If this is not the situation, but instead a choice about several decision alternatives evaluated through economic criteria (e.g., a comparison of alternative data quality enhancement procedures) is made based on the metric values, then requirements (R1) and (R2) are of significant importance.

R3: Quality of The Setup Settings And The Assessment Of The Metric Values

(R3) seeks to ensure that regardless of the measuring subjects, the intended measurement is accurately achieved. Consequently, the necessity for validity, reliability, and objectivity is often of significant relevance, as exemplified by the evaluation of the data quality component of currency. In practical applications, internal validity holds significant importance. Internal validity primarily concerns if the defined currency ("object of interest") is accurately assessed by the metric. Secondly, it guarantees that substantial alterations in the metric values (i.e., the dependent variable) are genuinely attributable to modifications in the variables affecting currency, rather than extraneous influences (control variables). Conversely, external validity is predominantly significant when the measure is applied just to a sample of the dataset, although the findings are utilised to make inferences about the entire dataset. Reliability seeks to ensure that the metric produces consistent or nearly same findings (i.e., high stability of outcomes) in repeated evaluations of the same data (e.g., throughout time), hence guaranteeing accurate measurement in this context. Objectivity is essential for facilitating automated data quality assessments and acquiring metric values that are unaffected by external factors, such as varying interviewers.

Inadequate data quality metrics (R3) may yield inadequate metric values (cf. above). Concerning the economically driven management of data quality, it is particularly challenging to assess the data quality level before to and subsequent to implementing a data quality enhancement initiative. A measure that does not satisfy (R3) cannot provide reliable conclusions regarding the actual change in data quality levels. Consequently, a data quality

enhancement initiative may be seen beneficial yet may not genuinely enhance data quality or may do so only little.

In summary, the following considerations must be addressed while formulating and implementing a metric:

It is essential to evaluate the necessary data values, metadata, and parameter values for the instantiation and application of a data quality metric. When substantial historical data, whether from internal or external sources, including big or open data, is accessible, the requisite data values and parameters, particularly the configuration parameters, can be ascertained in a valid, objective, and reliable manner through statistical methods. If such a data foundation is unavailable, expert estimations are required, which must also be acquired transparently and verifiably.

(b) Metrics should be explicitly stated to guarantee that, if the necessary data values and parameters are properly specified, the calculation rule ensures (R3), particularly objectivity and reliability. If the calculation rule cannot be officially articulated, the computation of the metric values must be delineated in a systematic, transparent manner and as clearly as possible to facilitate intersubjective application. It is essential to guarantee the alignment between the intended measurement (namely, a precise specification of the relevant data quality dimension) and the actual measurement (operationalisation of the stated data quality dimension).

R4: Robust Compilation Of The Metric Values

(R4) has significant importance when the evaluation or selection of decision alternatives is not just reliant on the isolated data quality assessment of an individual data point. Specifically, let us examine an application designed just to assess the completeness of data values for a characteristic, independent of one another. Utilise the individual metric values directly for decision-making; for instance, if no data value (or a value semantically equal to 'NULL') is present, execute action a; otherwise, refrain from taking any action. An isolated judgement is made based on the data value levels, necessitating no aggregation. Nevertheless, practical choices, such as the implementation of data quality enhancement initiatives, are typically not founded solely on a singular data point or individual data values assessed in isolation. This criterion is particularly pertinent in several decision-making scenarios that depend on the data quality of extensive datasets. The quality of data within a substantial portion of a customer database, or the entire database, may be evaluated to determine the feasibility of initiating a marketing campaign.

In summary, if a metric was not specifically formulated for assessing the data quality of individual data values (or is not exclusively utilised in such contexts), but instead is intended or employed to convey the data quality of multiple data values within a singular metric value, (R4) is especially pertinent. The greater the significance of this aggregated measure value for decision-making, the more pertinent (R4) becomes.

R5: Economic Efficacy Of The Metric

(R5) is of paramount importance if evaluating data quality using a metric incurs significant expenses. The metric values are utilised for a choice with possibly significant costs and advantages. Given that inadequate data quality frequently incurs significant expenses in practice, this need must be considered throughout the metric design phase. Specifically, let us examine an application designed to assess the completeness of the data values for a singular property within a relation of around 100 tuples. The evaluation is performed manually by one individual during a duration of five minutes, rendering the expenses for ascertaining both the configuration settings and the metric values insignificant. This individual archives the assessment outcome (i.e., the percentage of complete data values according to the metric) solely for documentation, without conducting further analysis or making decisions based on the assessment result (i.e., any potential benefits from applying the metric are inconsequential). In such instances, assessing the efficacy of the metric, especially in relation to other metrics that may provide marginally quicker counting, is often unnecessary. One may contend that assessing the efficacy of metrics is unnecessary in the context of a data quality evaluation mandated by legal laws (e.g., in risk management). One may similarly argue that assessing efficiency is not pertinent to the choice about the use of a measure. Nonetheless, this reasoning may be inadequate: Even with a compulsory review, a corporation may reassess the economic efficacy of many criteria to choose the most suitable option. Consequently, in several instances, (R5) holds significant practical relevance. Furthermore, (R5) holds significant relevance in evaluating data quality within the context of data governance or data quality management initiatives, which are often focused on economic efficiency.

In summary, data quality measures are often not intended for evaluating data quality in situations when the economic significance of the evaluation is minimal, rendering both the potential benefits and the associated costs inconsequential. Consequently, the significance of (R5) is evident. This importance escalates with the anticipated expenses, respectively. The anticipated advantages of both evaluating data quality and the judgements derived from the evaluation include.

Conclusion

This study presents five needs for data quality measurements to facilitate decision-making under uncertainty and economically focused data quality management. Our needs enhance the current literature in two respects. In contrast to current methodologies that are disjointed and ambiguous, we propose a comprehensive set of well stated standards, facilitating straightforward and transparent verification. This is crucial for practical applications. Secondly, unlike existing works, we substantiate our criteria using a robust decision-oriented framework. In the absence of such a framework, it is impossible to validate the significance of the needs, nor is it evident what consequences arise from a failure to meet a requirement. Consequently, our needs are crucial for assessing current metrics and for developing new metrics, particularly within the framework of Design Science Research. Inadequate measurements, which may result in erroneous judgements and financial losses, can be

detected and enhanced based on our criteria. The relevance and effectiveness of the suggested rules are illustrated using five established data quality measures. Both outcomes are essential from both a methodological and practical perspective.

References

[1] Malhotra, I., Gopinath, S., Janga, K. C., Greenberg, S., Sharma, S. K., & Tarkovsky, R. (2014). Unpredictable nature of tolvaptan in treatment of hypervolemic hyponatremia: case review on role of vaptans. Case reports in endocrinology, 2014(1), 807054.

[2] Gonugunta, K.C. and K. Leo. (2018) Oracle Analytics to Predicting Prison Violence. International Journal of Modern Computing. 1(1): 23-31.

[3] Singh, V. K., Mishra, A., Gupta, K. K., Misra, R., & Patel, M. L. (2015). Reduction of microalbuminuria in type-2 diabetes mellitus with angiotensin-converting enzyme inhibitor alone and with cilnidipine. Indian Journal of Nephrology, 25(6), 334-339.

[4] Karakolias, S. E., & Polyzos, N. M. (2014). The newly established unified healthcare fund (EOPYY): current situation and proposed structural changes, towards an upgraded model of primary health care, in Greece. Health, 2014.

[5] Gonugunta, K.C. (2018) Role of Analytics in Offender Management Systems. The Computertech. 27-36.

[6] Shilpa, Lalitha, Prakash, A., & Rao, S. (2009). BFHI in a tertiary care hospital: Does being Baby friendly affect lactation success?. The Indian Journal of Pediatrics, 76, 655-657.

[7] Pasham, S.D. (2018) Dynamic Resource Provisioning in Cloud Environments Using Predictive Analytics. The Computertech. 1-28.

[8] Polyzos, N. (2015). Current and future insight into human resources for health in Greece. Open Journal of Social Sciences, 3(05), 5.

[9] Gonugunta, K.C. (2018) ZDL-Zero Data Loss Appliance–How It Helped DOC in Future-Proofing Data. International Journal of Modern Computing. 1(1): 32-37.

[10] Gopinath, S., Janga, K. C., Greenberg, S., & Sharma, S. K. (2013). Tolvaptan in the treatment of acute hyponatremia associated with acute kidney injury. Case reports in nephrology, 2013(1), 801575.

[11] Gopinath, S., Giambarberi, L., Patil, S., & Chamberlain, R. S. (2016). Characteristics and survival of patients with eccrine carcinoma: a cohort study. Journal of the American Academy of Dermatology, 75(1), 215-217.

[12] Gonugunta, K.C. and K. Leo. (2017) Role-Based Access Privileges in a Complex Hierarchical Setup. The Computertech. 25-30.

[13] Swarnagowri, B. N., & Gopinath, S. (2013). Ambiguity in diagnosing esthesioneuroblastoma--a case report. Journal of Evolution of Medical and Dental Sciences, 2(43), 8251-8255.

[14] Gopinath, S., Ishak, A., Dhawan, N., Poudel, S., Shrestha, P. S., Singh, P., ... & Michel, G. (2022). Characteristics of COVID-19 breakthrough infections among vaccinated individuals and associated risk factors: A systematic review. Tropical medicine and infectious disease, 7(5), 81.

[15] Shilpa, Lalitha, Prakash, A., & Rao, S. (2009). BFHI in a tertiary care hospital: Does

being Baby friendly affect lactation success?. The Indian Journal of Pediatrics, 76, 655-657.

[16] Gopinath, S., Janga, K. C., Greenberg, S., & Sharma, S. K. (2013). Tolvaptan in the treatment of acute hyponatremia associated with acute kidney injury. Case reports in nephrology, 2013(1), 801575.

[17] Gopinath, S., Giambarberi, L., Patil, S., & Chamberlain, R. S. (2016). Characteristics and survival of patients with eccrine carcinoma: a cohort study. Journal of the American Academy of Dermatology, 75(1), 215-217.

[18] Gonugunta, K.C. (2016) Oracle performance: Automatic Database Diagnostic Monitoring. The Computertech. 1-4.

[19] Pasham, S.D. (2017) AI-Driven Cloud Cost Optimization for Small and Medium Enterprises (SMEs). The Computertech. 1-24.

[20] Gonugunta, K.C. (2018) Apply Machine Learning Oracle Analytics–Combined. The Computertech. 37-44.