Research on Data Quality Assessment Techniques Based on Data Governance Frameworks

Authors: Li Yuxuan

Date: 2025-05-15T00:00:00+00:00

Abstract

Currently, data quality issues constitute the criterion for determining the efficiency of data-driven decision-making, business operations, and the effectiveness of policies and institutions. However, most existing data quality assessment methods are unable to adapt to increasingly complex environments. By leveraging comprehensive data governance techniques and integrating time series models with regression models, this study establishes a novel quantitative assessment methodology for data quality evaluation, thereby improving the precision of assessment outcomes. Finally, through a synthesis of theoretical and practical research findings, this paper discusses the practical significance of the research topic "Data Quality Assessment Technology Based on Data Governance Framework" in the big data era, addressing current challenges and future development prospects. Employing new quantitative evaluation methods to resolve certain issues emerging in big data environments, this approach facilitates the proper direction of data governance development and enables more effective application within data governance practices.

Full Text

Preamble

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065

Abstract

Data quality issues are critical determinants of decision-making efficiency, business operations, and policy effectiveness in the modern era. However, most existing data quality assessment methods are ill-equipped to handle increasingly complex environments. This study leverages comprehensive data governance techniques to integrate time-series models with regression analysis, creating a

novel quantitative evaluation methodology for data quality assessment that significantly improves the precision of evaluation results. By combining theoretical and practical research findings, this paper discusses the practical significance of the research topic "Data Quality Assessment Technology Based on the Data Governance Framework" in the big data era, addressing current challenges and future development prospects. The proposed quantitative evaluation approach helps resolve emerging issues in big data environments, guides data governance toward proper development, and enables more effective applications in data governance practice.

Keywords: Data Quality Assessment, Time-series Model, Regression Analysis, Data Governance Framework

1. Introduction

Data quality issues such as noise, missing values, and inconsistencies occur throughout the entire data lifecycle, including collection, transmission, storage, and application. Data governance serves as a crucial mechanism for ensuring data quality and enhancing data value. A primary task of data governance involves evaluating data quality, where appropriate assessment methods provide valuable references for governance initiatives and help improve data quality issues, ultimately leading to the formulation of superior data management systems.

The concept of "data governance" originated from Watson et al.' s 2004 research on "data warehouses," which laid the foundation for subsequent studies in this domain. Wang Nana, in the work "WH Company Data Governance Planning Based on Big Data Platforms," defines data governance as the process of unified planning, standardization, and rule-making for scattered and disorganized data. Data governance represents a prerequisite for effective operations on information data and a powerful guarantee for improving data quality, making it a long-standing focus of attention. It primarily encompasses data management, standardization, security, and quality supervision and evaluation. Based on these elements, data governance completes relevant analysis and improvement adjustments for data quality. Through scientific assessment of data quality aspects, it enables effective identification of quality issues and facilitates cleanup and repair operations on problematic data.

Data governance typically follows a four-layer structural model comprising the strategic layer, tactical layer, execution layer, and operations/support layer. From a subject perspective, data governance primarily involves various elements under organizational internal control, focusing on data within the organization while emphasizing the application of big data management technologies by organizational personnel.

1.1 Research Background

With the development of big data and artificial intelligence, data applications have become increasingly widespread across industries. Ensuring data quality is key to scientific research, business decision-making, and realizing data value. However, traditional data quality assessment methods suffer from drawbacks such as single evaluation dimensions and difficulty meeting big data requirements. Therefore, there is an urgent need to seek new assessment approaches and technologies to address these emerging demands.

1.2 Research Significance

This research innovates data quality assessment practices to overcome the limitation of existing methods in handling dynamic environments. Most current assessment methods remain at a static level, making it difficult to reflect how data quality should be evaluated under dynamic conditions. By applying time-series models and regression analysis, we can assess data quality from both temporal and inter-relationship perspectives, making evaluations more precise. This approach compensates for deficiencies in existing assessment methods to some extent.

The research results provide a scientific assessment methodology for enterprises, governments, and research institutions, helping improve data quality management and thereby enhancing decision-making efficiency and business competitiveness. In complex big data application scenarios, industries such as finance, e-commerce, and healthcare face numerous data quality issues with associated risks and costs. This study' s findings can provide more accurate data quality assessment for these industries. Improving data quality not only yields direct economic benefits for enterprises and research institutions but also optimizes social resource allocation and enhances public service delivery. Providing reliable, high-quality data for government data management and public information services enables better service to economic and social sustainable development.

2. Research Status Introduction

This section reviews the current state of research in data governance frameworks, time-series models, regression analysis, and data quality assessment.

2.1 Basic Concepts

Data Governance

Data governance ensures data correctness, appropriate sharing, and security. A robust data governance approach can generate benefits for enterprises through optimized decision-making, cost savings, risk control, and enhanced security compliance, thereby increasing revenue and profits. Current research on data governance includes works by Zhu Wujin (2020) and Zeng Fan et al. (2021).

Internationally, scholars began exploring data governance much earlier. For instance, Donaldson A (2004) understood data governance from a compliance and regulatory perspective as a series of behaviors combining numerous rules and regulations. Thomas G and Putro BL (2006) viewed data governance from a responsibility perspective as organizational decision-making about data assets and responsibility allocation.

Data Governance Framework

Song Feng, in the book "Research on Intelligent Electric Drive Bridge Forward-Looking Technology Application Based on DMBOK," points out that DMBOK treats data as a special form of asset—distinct from other assets in that data is not consumed during use and can be utilized infinitely. To obtain the benefits of data governance requires big data, excellent data analysis tools, and the discovery of data quality issues with proposed solutions to help enterprises reduce business risks from unqualified data quality and improve data utilization value and decision support.

The data governance conceptual architecture is illustrated in [Figure 1: see original paper].

Time-Series Models

Li Lingling notes in "Application of Time-Series Model ARIMA in Data Analysis"that a time series refers to a sequence of data points arranged in chronological order, repeating at fixed or irregular intervals. Data includes three types: timestamp, period, and interval. The ARIMA model, or Autoregressive Integrated Moving Average model, is a widely used statistical model for time-series data.

Many models are included in this family, such as AR, MA, ARMA, and ARIMA models. For example, the AR model uses a linear combination of past values to explain the current state, essentially performing multiple regression across different time points within a period. An AR(p) model of order p can be expressed as:

$$X_t = \phi_0 + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \epsilon_t \quad (2-1)$$

The result is obtained by assigning different weights to past values of the time series and adding them together, combined with a constant term and random component.

The MA (Moving Average) model is defined as:

$$X_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q} \quad (2-2)$$

where θ_i are moving average coefficients, ϵ_t is the current period error term, and ϵ_{t-i} are random disturbances from previous periods.

The ARMA model combines AR and MA components. ARIMA includes an additional differencing term d. By differencing non-stationary time series, we obtain stationary sequences suitable for ARIMA. Its parameters are p, d, q, where d is the differencing order, denoted as ARIMA(p, d, q). The difference operations are expressed as:

First-order differencing:

$$\nabla X_t = X_t - X_{t-1} \quad (2-4)$$

Second-order differencing:

$$\nabla^2 X_t = \nabla(\nabla X_t) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \quad (2-5)$$

Higher-order differencing follows this pattern.

2.2 Domestic and International Research Status

Data Governance Research Status

The concept of data governance has evolved to establish domain-specific frameworks, including the DGI data governance framework [8], IBM data governance framework [Figure 2: see original paper], DAMA (Data Management Association) data governance framework [14] [FIGURE:3 and FIGURE:4], and China's "Data Governance Specification" national standard system.

Data governance has gained industry attention and sparked academic discussion. International scholars such as Nofie, Anonymous, and Paschal [15-17] have contributed to the ethical foundation of data governance frameworks through literature review across academic databases. Ronak Pansara [8] addressed inconsistencies between data repositories by establishing unified standards. Wang et al. [9] designed a big data healthcare platform with multi-source heterogeneous integration for massive high-dimensional data governance in large hospitals.

Domestic research on data governance started later, often directly adopting foreign concepts with superficial theoretical exploration of framework elements. However, Wu Xindong et al. proposed the HACE theorem in 2014, describing big data processing framework characteristics from three perspectives: big data processing, applications, and mining [Figure 5: see original paper]. National data security governance top-level design constitutes the national data strategy. Major countries and international organizations have issued strategic documents to guide digital economy development and data security governance [Figure 6: see original paper].

Data Quality Management Research Status

International academia and industry have conducted extensive research and practice on data quality management. Redman (2020) treats data quality as

a critical asset of the information age, exploring excellence in data collection and use [21]. Batini and Scannapieco (2021) provide in-depth investigation of data quality concepts, methodologies, and techniques [22]. Wang and Strong (2020) emphasize that data quality importance extends beyond accuracy to its meaning for data consumers [23]. Pipino et al. (2020) offer a comprehensive perspective on data quality assessment [24]. Enterprises and research institutions have developed comprehensive data quality management systems such as IBM InfoSphere and Microsoft SQL Data Quality Services. Wang and Reddy (2020) established a systematic framework for data quality assessment [26], while Hu and Zhao (2021) surveyed data quality issues in big data [27].

Domestic research is gradually developing, with some universities and enterprises achieving results in data governance and quality assessment. Li Xiangyang and Zhou Xiaofang (2020) elaborated on data governance concepts, frameworks, and practices [28]. Chen Gang and Zhao Yu (2022) discussed data quality and governance issues combining theory and practice [29]. Liu Qing and Zhang Xiaolin (2021) explored big data quality and data governance [30]. While progress has been made domestically and internationally, domestic research still lags behind in assessment precision, diversity, and automation, requiring continuous strengthening.

3. Current Problems

Data Quality Inconsistency

While "data assetization" has become common consensus, not all data qualifies as data assets—much of it constitutes junk data. True data governance should focus on valuable data assets that generate value, rather than indiscriminately managing all data.

Data Exchange and Sharing Challenges

Due to the lack of overall information planning at the outset, enterprise informatization has produced many independently developed, business-unit-driven monolithic systems and packaged software with different architectures, development languages, and databases. This has created numerous "information silos" where data remains isolated, preventing meaningful interoperability and hindering data from fulfilling its potential. Only by breaking down these silos and establishing data connectivity can data drive business development and management transformation, fully releasing big data value.

Lack of Effective Management Mechanisms

While most enterprises recognize data importance and attempt to control data mobility through business processes, practical implementation lacks effective management measures. Subjective factors cause issues during data flow, including maintenance errors, duplication, inconsistency, and incompleteness, generat-

ing massive amounts of junk data. This leads to unclear data ownership, chaotic management responsibility relationships, and ambiguous usage management, ultimately resulting in poor data quality and distortion.

Data Security Risks

Data security incidents have occurred frequently in recent years. In March 2018, Facebook's user data leakage and abuse caused a 7% stock price drop and $36 billion market value loss, with Cambridge Analytica declaring bankruptcy. China has also experienced similar issues: in 2011, CSDN' s user database was publicly exposed with over 6 million plaintext email accounts and passwords; in 2016, a SF Express employee stole tens of thousands of customer personal information; in 2017, a JD.com employee stole nearly 5 billion user information records for dark web sale. Data asset management is transitioning from decentralized manual management to centralized computer-based information management, with increased emphasis on data security.

4. Further Work

Dynamic Data Quality Assessment Challenges

Data quality is affected by various factors throughout the entire process of collection, transmission, storage, and application. Quality conditions change dynamically, necessitating real-time monitoring and evaluation mechanisms to identify problems promptly and enable correct decisions. Such mechanisms must track data quality changes, detect issues during data processing and computation, and provide real-time feedback to support decision-making. This requires assessment methods with strong flexibility and adaptability to handle gradually changing data and large-scale datasets. [Figure 7: see original paper] illustrates traditional static data governance access data analysis strategies.

Data Quality Assessment Standardization Challenges

To facilitate comparison and analysis of data from various sources, unified evaluation standards are needed. However, establishing universal standards applicable to diverse data sources is challenging due to data heterogeneity and complexity. Standards must consider multiple dimensions including accuracy, completeness, consistency, reliability, and timeliness. Additionally, unified standards for assessment methods, tools, and processes are needed to ensure consistency across all stages, enabling results to be compared at the same level. Achieving these goals can significantly improve assessment efficiency and effectiveness, benefiting data sharing and utilization.

Application of Mathematical Models in Data Quality Assessment

Mathematical models, including time-series and regression analysis, provide scientific methods for data quality evaluation. They enable quantitative analysis

of quality trends, identify relationships between quality and influencing factors, and predict future quality developments. However, efficiently applying such models to large-scale, multi-dimensional heterogeneous data remains an unresolved challenge. This section analyzes issues of model selection, parameter setting, data preprocessing, and result interpretation. Solving these problems can improve assessment accuracy and scientific rigor, providing stronger support for data governance.

4.3 Considered Mathematical Models

This study employs comprehensive research methods to explore important issues in data quality evaluation, utilizing statistical and mathematical knowledge to construct quantitative analysis models. The models primarily include time-series models and regression analysis models, which accurately reflect data quality changes and explain influencing factors through a series of variables. Time-series models reveal how various data quality indicators change over time, while regression analysis identifies key influencing factors and their impact degrees.

To meet prediction accuracy and real-time strategy update requirements, this research further explores introducing deep learning and reinforcement learning algorithms into data quality assessment. Both belong to machine learning and possess strong capabilities for handling complex data patterns, providing solutions for dynamic data environments.

The study combines simulation experiments and case study methods to test and analyze algorithms and strategies across different data and application scenarios. Simulation experiments use simulators to replicate various network conditions and data flows for model effectiveness examination, while case studies investigate model applications and limitations through real business scenarios and datasets.

As previously defined, the AR (Autoregressive) model takes the form:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

where X_t is the current data quality indicator and X_{t-i} are past indicators. This model captures the relationship between current and historical data quality metrics. Model order can be determined through ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function).

ACF describes how autocorrelation of the data quality indicator time series changes with lag order. If ACF shows a tailing-off pattern (autocorrelation coefficients gradually decreasing but not reaching zero) and PACF exhibits cut-off after a certain lag (coefficients suddenly becoming zero or near-zero), this typifies an AR model.

In dynamic data quality assessment, AR models capture time-varying characteristics. For instance, if data quality at a given moment depends on previous equipment states, an appropriately ordered AR model can reflect this relationship. Similarly, for data accuracy metrics, low accuracy at one moment may lead to low accuracy at the next, which AR models can learn from historical correlations to enable real-time quality change monitoring.

MA models are more troublesome for order determination, typically relying on ACF cutoff lag, but become difficult to judge under big data volume, multi-dimensionality, and inconsistency. I models struggle with multi-dimensional assessment because differenced data quality indicators across dimensions may vary significantly and cannot be compared within the same framework like AR models.

Therefore, analyzing ACF and PACF patterns reveals that AR models offer certain advantages over MA and I models in addressing dynamic assessment, standardization, and mathematical model application issues in data quality evaluation.

References

[1] Wang Nana. Data Governance Planning for WH Company Based on Big Data Platform[D]. Jilin University, 2021. DOI:10.27162/d.cnki.gjlin.

[2] AL-RUITHE M, BENKHELIFA E, HAMEED K. A Systematic Literature Review Governance Cloud Data Governance[J]. Personal and Ubiquitous Computing, 2019, 23(5-6): 839−859.

[3] Cheng Ping, Chang Ji, Xia Hui. Accounting Big Data: Connotation, Framework, and Technical Implementation[J]. Commercial Accounting, 2022, (12): 4-9.

[4] He Yuan. Data Law[M]. Peking University Press, 2020.

[5] Chen Hu, Chen Jian. Accounting Big Data Analysis and Processing Technology: Boosting Data Empowerment for Finance' s New Future, 2022, (10): 23-2.

[6] Song Feng, Li Jiakuo, Wei Yongxiang, et al. Research on Intelligent Electric Drive Bridge Forward-Looking Technology Application Based on DMBOK[J]. Mechanical & Electrical Engineering Technology, 2023, 52(05): 159-162+227.

[7] Hu Benli. "DAMA Data Management Body of Knowledge (2nd Edition)"New Book Subscription Preview[J]. Project Management Technology, 2020, 18(08): 143.

[8] Ekundayo Bhaumik Chinoperekweyi J. Identifying the core data governance framework principle: a framework comparative analysis[J]. Organization Leadership Development Quarterly, 2023, 5(1): 30-53.

[9] Wang M, Li S, Zheng T, et al. Big data health care platform with multisource heterogeneous integration massive high-dimensional data governance large hospitals: design, development, application[J]. JMIR Medical Informatics, 2022, 10(4): e36481.

[10] Li Lingling, Xin Hao. Application of Time-Series Model ARIMA in Data Analysis[J]. Fujian Computer, 2024, 40(04): 25-29.

[11] Li Yongdi. GDP Volume and Growth Rate Prediction Based on ARIMA Deflator Method and Principal Component Regression Model—Taking Henan Province as an Example. Henan Science and Technology, 2021, 40(13): 149-155.

[12] Wang Yingwei, Ma Shucai. Time Series Prediction Based on ARIMA and LSTM Hybrid Model. Computer Applications and Software, 2021, 38(2): 291-298.

[13] Majumder S, Bhattacharjee A, Kozhaya J N. Enhancing AI Governance in Financial Industry through watsonx.governance[J]. Authorea Preprints, 2023.

[14] Wang M, Li S, Zheng T, et al. Big data health care platform with multisource heterogeneous integration massive high-dimensional data governance large hospitals: design, development, application[J]. JMIR Medical Informatics, 2022, 10(4): e36481.

[15] Nofie I. The fight for our personal data: analyzing the economics of privacy digital platforms[J]. International Journal of Management, 2024, 66(6): 774-791.

[16] Anonymous. Modern Integration and Governance for the AI Era[J]. Database Trends and Applications, 2024, 38(5): 20.

[17] Paschal O, Damian E, Carsten B. Corrigendum: Towards understanding of global brain data governance: ethical positions that underpin global brain governance discourse[J]. Frontiers Data, 2023, 61344345-1344345.

[18] Pansara R. Review & Analysis of Master Data Management in Agtech Manufacturing industry[J]. International Journal of Sustainable Development in Computing Science, 2023, 5(3): 51-59.

[19] Tan Zhanglu, Wang Meijun. Research on Conceptual Model and Technical Architecture of Intelligent Coal Mine Data Governance[J]. Journal of Mining Science, 2023, 8(2).

[20] Deng Junzeng. Discussion on Hospital Health Medical Data Governance[J]. Journal of Medical Informatics, 2021, 42(8): 14-17.

[21] Redman, T. C. (2020). Data Quality for the Information Age: How to Achieve Excellence in the Information You Collect and Use. Morgan Kaufmann.

[22] Batini, C., & Scannapieco, M. (2021). Data Quality: Concepts, Methodologies, and Techniques. Springer.

[23] Wang, R. Y., & Strong, D. M. (2020). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 37(1), 47-81.

[24] Pipino, L., Lee, Y. W., & Wang, R. Y. (2020). Data Quality Assessment. Communications of the ACM, 63(1).

[25] Delone, W. H., & McLean, E. R. (2021). The DeLone and McLean framework for information systems success: Where we are, and where are going. Journal of Association Information Systems, 22(1), 53-94.

[26] Wang, R. Y., & Reddy, M. P. (2020). A Taxonomical Study of Data Quality Assessment. IEEE Transactions on Knowledge and Data Engineering, 32(8).

[27] Hu, Q., & Zhao, J. L. (2021). Data Quality in Big Data: A Survey. IEEE Access, 9, 22226-22246.

[28] Li Xiangyang, & Zhou Xiaofang. (2020). Data Governance: Concepts, Frameworks, and Practices. Electronic Industry Press.

[29] Chen Gang, & Zhao Yu. (2022). Data Quality and Data Governance: Theory and Practice. Economic Management Press.

[30] Liu Qing, & Zhang Xiaolin. (2021). Big Data Quality and Data Governance. Computer Research and Development, 58(1), 1-15.

Note: Figure translations are in progress. See original paper for figures.