Available Online: 23 October 2026 DOI: 10.54254/3029-0880/2026.38151

Technical distortions of statistical data in the big-data era: causes and countermeasures

Quwen Qian

Linkou Senior High School, Suzhou, China

qianquwen@example.com

Abstract. The growing application of big-data technologies has made official statistics a critical basis for macroeconomic judgment and public policy formulation. Existing scholarly work mostly concentrates on deliberate data falsification induced by administrative interventions and performance-oriented incentives, while systematic discussions about unintentional distortions stemming from technical factors remain fragmented. This paper distinguishes technical statistical distortion from intentional human data manipulation. Based on literature-driven qualitative analysis, this paper sorts out typical forms of technical distortion amid big-data scenarios, and identifies four major technical root causes: flawed statistical planning, outdated data collection and input procedures, non-standard computational algorithms and insecure distributed data storage. The paper proposes targeted preventive strategies in four areas: source-end scheme optimization, intelligent audit system upgrade, standardized data processing routines and traceable storage architecture construction. Research outcomes indicate that institutional adjustment alone cannot completely eliminate technical distortion; multilevel technical governance must span the full data lifecycle to reduce systematic bias in statistical outputs. This study offers practical references for enhancing data quality under large-scale multi-source data environments.

Keywords: big data, statistical data quality, technical distortion, total survey error, data lifecycle management

1. Introduction

The quality of statistical data directly determines the credibility of economic monitoring results, resource allocation decisions and government policy arrangements. Within today's big-data context, statistical departments combine outputs from traditional questionnaire surveys with massive heterogeneous multi-source information sourced from third-party platforms, administrative archives and sensor devices. Such integration greatly enriches available data resources, yet simultaneously brings about new types of potential data errors [1].

Chinese academia has long examined the problem of statistical data distortion. A large body of existing research focuses on institutional drivers, such as assessment-oriented mechanisms, administrative interference and interest-driven intentional data tampering. In contrast, analyses on non-subjective technical deviations are mostly scattered across different research paragraphs, rather than being developed as independent, complete research themes [1]. Different from deliberate falsification, technical distortion arises from workflow

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

loopholes, tool constraints and methodological defects instead of subjective fraud behaviours [2]. This creates an obvious research gap: few studies specifically discuss how technical defects across each statistical procedure generate data deviations, particularly fresh risks brought by multi-source data fusion under big-data conditions [3, 4].

Unlike previous research that combines institutional and technical distortion factors in a single analytical framework, this paper excludes intentional data falsification and focuses on unintentional technical distortions arising across the full statistical workflow: statistical scheme design, data acquisition, manual entry, logical audit, data storage and numerical computation [5]. It sorts out relevant conceptual definitions, typical performance patterns, underlying technical triggers and corresponding countermeasures.

Its practical value lies in reminding statistical practitioners that high-quality statistical outcomes rely not only on sound institutional arrangements, but also on rigorous technical governance across every stage of the data lifecycle [1]. The conclusions can also serve as learning materials for high-school and undergraduate statistics learners to understand real-world error sources in official statistical practice. The main conclusions are elaborated in later sections. The following text first defines the concept and manifestations of technical distortion; second, it analyzes the technical causes of distortion throughout the data lifecycle; and finally, it proposes preventive technical governance measures for the entire process.

2. Connotation and typical manifestations of statistical data technical distortion

2.1. Conceptual distinction: technical distortion versus deliberate falsification

It is essential to distinguish clearly between technical distortion and intentional data falsification. Deliberate falsification is a subjective human act: relevant personnel alter statistical figures to cope with performance evaluation requirements, pursue personal interests or respond to administrative pressure, with full awareness that they are producing incorrect statistical results.

On the contrary, technical distortion represents unintended biases and deviations produced by defective operation procedures, improper tool selection, methodological shortcomings or system-level limitations [2]. Statistical practitioners have no intention to distort real-world facts; errors emerge naturally during daily operational steps. Even when all relevant staff strictly abide by professional ethical codes, technical distortion may still take place. From a statistical perspective, it falls into the category of systematic and random errors rather than fraud. In practice, a single dataset may contain both intentional tampering and technical distortions; the two are easily confused with one another, which is the practical reason why past studies have tended to analyze these two types of issues as a single category [1]. This paper sets intentional falsification aside and concentrates solely on technical-origin deviations.

2.2. Traditional manifestations of technical distortion

Prior to the widespread application of big-data technologies, technical distortion presented several recurring patterns within conventional statistical workflows. First, inconsistent comprehension of statistical indicators: investigators or respondents interpret survey indicators differently due to ambiguous conceptual definitions. Second, systematic sampling bias: outdated sampling frames, insufficient sample size, or poorly designed stratification lead to samples that fail to reflect the characteristics of the overall population. Third, human-caused input mistakes: paper-based forms, oral reports and manual data transcription produce typographical errors, missing entries and duplicated records [2]. Fourth, computational deviation: improper

application of statistical formulas including average calculation, year-on-year and month-on-month comparison and weight allocation generates miscalculated outputs. Fifth, local storage risks: files stored in local devices may get modified accidentally without retaining complete change logs [3]. When data quality issues are entirely attributed to human fraud, governance measures tend to focus excessively on institutional accountability, while overlooking optimizations at the process and tool levels; as a result, the root causes of technical errors remain unaddressed [1].

2.3. New forms of technical distortion under big-data background

Big-data applications bring new error patterns rarely seen in conventional small-sample surveys. The first prominent challenge is inconsistent statistical standards among heterogeneous multi-source datasets. Different government departments, commercial platforms and social organisations adopt inconsistent indicator definitions, statistical coverage and statistical cycles. Without unified metadata management mechanisms, direct fusion of these datasets will create hidden deviations in integrated results [4].

Second, bias arises during the conversion of heterogeneous data. When researchers convert unstructured texts, transaction logs and sensor-captured information into structured statistical tables, format conversion algorithms may discard valid information or introduce extra noise [5]. Third, failure to identify abnormal values amid massive datasets: traditional manual inspection and simple filtering rules become ineffective when handling millions of data entries. Outliers and low-quality data may pass inspection without triggering warning signals and contaminate the whole dataset [2]. Compared to distortions in traditional statistical techniques, sources of distortion in a big data environment have expanded from manual operations to include multi-source integration and algorithmic transformations; errors are more difficult to detect, making manual troubleshooting significantly more challenging [4].

3. Core technical causes behind technical distortion of statistical data

Root causes can be grouped into four stages across the data lifecycle: defects in statistical scheme design, outdated data collection and input methods, nonstandard data processing approaches, and vulnerabilities in data storage systems [2]. Each section covers both classic statistical risks and unique challenges brought by big-data development [4].

3.1. Shortcomings in statistical design techniques

Statistical scheme design acts as the starting point of data production. Defects formed at this phase will spread to downstream procedures and turn into inherent systematic bias [1].

Firstly, defects in questionnaire and indicator design: vague indicator definitions, inconsistent statistical standards, misleading survey questions and missing key dimensional settings. Without unified filling specifications, respondents and data coders will produce inconsistent outputs even when facing identical objective facts. Secondly, non-standard sampling arrangements—obsolete sampling frames, biased sample selection, insufficient sample size, and unreasonable stratified rules—will trigger systematic sampling error and weaken sample representativeness for target populations [2].

Big-data scenarios pose an additional challenge: multi-source datasets often lack unified metadata specifications. Metadata records indicator definitions, coverage, time ranges, and original data sources. Without unified metadata tools, cross-departmental data cannot be accurately aligned during integration, which further generates fusion-related deviations [4]. Many institutions directly combine external big-data resources without sorting metadata information, which constitutes a major source of hidden technical

distortion [5]. Design-stage problems are particularly tricky because later-phase processing can hardly fix flaws that exist from the very beginning [1].

3.2. Outdated data collection and manual input technologies

Manual-centred collection modes remain prevalent in grassroots statistical work. Paper questionnaires and on-site interviews inevitably lead to missing data, misreporting, and duplicate submissions. Random human errors cannot be thoroughly eliminated no matter how carefully operators conduct their work [2].

Audit mechanisms represent another weak link. Many information systems lack built-in automatic logical verification rules, and audit work largely depends on manual reading and checking. Logical contradictions between different variables are likely to escape detection. Due to practical constraints such as funding limitations and delays in system upgrades, some grassroots units have seen slow progress in the intelligent transformation of their systems and must continue to rely on manual review processes [1].

Under big-data contexts, large volumes of unprocessed heterogeneous raw data flow into statistical pipelines without pre-filtering mechanisms. Low-quality data containing null values, outliers and repeated entries directly enter subsequent analytical procedures and continuously amplify initial errors [5]. Even sophisticated big-data algorithms cannot fully repair defects formed at the data collection stage [4].

3.3. Non-standard data-processing approaches

Distortions formed during data processing can be divided into two categories. The first results are from improper application of computational methods. Practitioners may select inappropriate averaging algorithms, set weight coefficients incorrectly or misunderstand year-on-year and chain comparison formulas, which eventually result in wrong calculation outcomes. Such mistakes stem from technical defects rather than subjective data fraud [2].

The second type arises from insufficient data-cleaning capacity. Without systematic cleaning specifications and weak outlier identification capabilities, abnormal observations remain in datasets and distort descriptive statistics and inferences [5]. One unique risk under big-data environments lies in the mechanical transplantation of general big data algorithms without adaptation to government statistical business requirements. Generic machine learning and data mining tools are developed for commercial-oriented scenarios. Direct application of such tools to official statistical work without targeted adjustment may violate basic statistical survey principles and produce unreliable analytical outputs [4]. Technical staff may blindly adopt popular big-data tools while ignoring basic statistical logic [1].

3.4. Technical vulnerabilities of data storage and management

Traditional local file storage faces risks of accidental data modification. With the wide adoption of distributed storage frameworks in big-data architecture, multiple nodes participate in data reading and writing operations. However, many business systems do not maintain a complete audit trail of data-writing, modification, and deletion operations [3]. When data changes, administrators are unable to accurately identify the user who performed the operation, the time of the operation, or the data status before and after the change. This lack of traceability not only magnifies technical discrepancies caused by unintentional errors but also makes it more difficult to identify malicious tampering [3].

4. Preventive technical countermeasures for technical distortion of statistical data

Corresponding countermeasures are proposed targeting each causal link above, covering the complete data lifecycle: source-end optimisation of statistical schemes, upgrading collection infrastructure together with intelligent audit technology, standardisation of processing workflows, and construction of traceable secure storage systems [1].

4.1. Optimise statistical design to reduce deviations at the source

First, optimise indicator and questionnaire design. Formulate explicit written definitions for each statistical indicator, unify statistical standards, coverage scopes and computational formulas, remove misleading questions and supplement missing dimensional settings. Release unified filling guidelines for investigators and survey respondents [2].

Second, implement scientific sampling arrangements. Update sampling frames on a regular basis, calculate sample size reasonably according to preset statistical error requirements, and standardise the implementation of stratified sampling and multi-stage sampling to minimise systematic sampling bias [1].

Third, develop unified metadata management tools for big-data multi-source integration. Metadata records indicator definitions, data sources, reporting cycles, and coverage. Before merging cross-departmental datasets, metadata comparison should be completed in advance to avoid blind data fusion caused by inconsistent standards [4]. This measure can effectively mitigate distortions triggered by multi-source heterogeneity at the design phase. Unified metadata management requires more than just development tools; it also requires cross-departmental collaboration mechanisms. Relying solely on technical tools cannot resolve conflicts in business definitions across different departments [5].

4.2. Upgrade data-collection infrastructure and deploy intelligent rule-engine auditing

Reduce over-reliance on pure manual paper-based reporting modes. Expand digital self-reporting channels and standardise data import interfaces for third-party external data. Embed multi-dimensional automatic verification rules in statistical information systems. Rule-engine-based intelligent early-warning systems can automatically check logical conflicts, value-range violations and abnormal numerical jumps among variables. Suspicious records are flagged for manual re-examination, reducing the omission risk of purely manual auditing [2].

For externally imported big-data resources, add pre-processing filtering gates before raw data enters formal statistical pipelines. Low-quality data including null values, extreme outliers and duplicated entries should be recognised and tagged at the collection gateway, instead of being transferred to later analytical modules [5]. Even intelligent systems cannot replace human judgment, so human review remains an essential supplementary step [1].

4.3. Standardise data-processing workflow and build multi-source cross-validation mechanisms

Statistical agencies and relevant institutions should compile standard operation manuals for data processing work. Manuals need to clarify applicable scenarios for average value calculation, weighted computation and index comparison algorithms to prevent method misuse [2]. Formulate formal data-cleaning specifications, and define judgment criteria as well as disposal rules for outlier observations [5, 6].

When utilising big-data analytical tools, practitioners need to adjust general-purpose algorithms to fit official statistical business logic instead of direct copy-and-paste application [4]. Establish a multi-source

cross-validation mechanism to compare and verify computational results from different sources, such as the platform, administrative records, and survey samples. When discrepancies arise in outputs from multiple sources, analysts should not arbitrarily favor the conclusions from a single data source; instead, they should first trace the metadata, data collection criteria, and data-cleaning rules to identify the root cause of the discrepancy. Big data should be used solely as a supplementary technical tool and cannot replace classical statistical theory and business rules [7].

4.4. Enhance storage traceability and adopt anti-tampering technical solutions

For distributed big-data storage platforms, strengthen operation-log management. Every writing, modification and deletion action should record operator identity, timestamp and change trails between original and updated values, ensuring full traceability of all data-related operations [3].

Drawing on previous research outcomes, blockchain-enabled technical mechanisms can be selectively introduced into key statistical-data storage modules. Hash values, timestamps and identity markers can record data status. Any alteration of stored records will cause obvious changes in hash values, making modifications detectable rapidly [3]. It provides technical evidence for tracing accidental operational errors or malicious tampering, and mitigates risks brought by distributed storage architectures. Meanwhile, statistical departments should cultivate interdisciplinary professionals equipped with both statistical theoretical knowledge and information-system capabilities [4, 8]. Advanced technical facilities cannot exert expected effects without competent operators. Good hardware and software provide only a foundation; competent personnel remain the decisive factor [1].

5. Conclusion

Research results show that statistical-data quality risks are not merely induced by institutional and human factors. Even when all practitioners comply with professional ethics, technical distortion may occur at every stage including design, collection, input, processing and storage. In big-data environments, multi-source heterogeneity, massive data volumes, and distributed storage introduce new hidden error sources. Improving statistical data quality requires joint efforts of institutional construction and systematic technical governance across the whole data lifecycle. To prevent statistical technical distortion, statistical agencies should optimise indicator-questionnaire design and develop unified metadata management tools to avoid blind multi-source data fusion at the source-end design stage. It is essential to reduce dependence on manual reporting, deploy intelligent rule-engine audit systems and set pre-processing filtering gates to block low-quality raw data before it flows into formal statistical workflows. Statistical agencies need to compile standard data-processing manuals and adapt general big-data algorithms to official statistical scenarios, and establish multi-source cross-validation mechanisms for result checking. For distributed storage systems, complete operation logs should be maintained and blockchain-enabled anti-tampering mechanisms may be adopted to realise full traceability of data operations, while cultivating interdisciplinary professionals to guarantee the implementation of technical measures. This research still has certain limitations. Adopting literature-based qualitative analysis, this paper does not carry out quantitative simulation or real-case empirical tests for each type of technical distortion. Follow-up research can conduct simulation experiments to quantify error amplitudes brought by different technical defects.

References

[1] Zhang, H. L. (2022). Research on quality control methods of statistical data under the background of big data. *Productivity Research, 2022*(5), 156-160.

[2] Liao, Y. H., & Chen, H. (2022). Analysis of influencing factors of statistical data quality under big data background. *Standard Technology, 2022*, 137-139.

[3] Yao, B., Wang, Z. S., Tian, A. X., et al. (2021). Research on government statistical data distortion based on blockchain technology. *Journal of Xianyang Normal University, 36*(6), 18-22.

[4] Yu, F. D. (2018). Application, bottleneck and integration path of big data in government statistics. *Survey World, 2018*(11), 3-11.

[5] Li, J. C. (2020). Quality control of big-data application. *Statistical Research, 37*(2), 119-128.

[6] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14*, 2-2.

[7] Kitchin, R. (2015). Big data and official statistics: Opportunities, challenges and risks. *Statistical Journal of the International Association of Official Statistics, 31*(3), 471-481.

[8] Schubert, P., Guiver, T., MacDonald, R., & Yu, F. (2006). Using quality measures to manage statistical risks in business surveys. In *European Conference on Quality in Survey Statistics*.