

Article

Automated Data Monitoring Using a Canberra-Based Drift Score

Konstantin Piryanov ^{1,*} , Iveta Grigorova ¹, Aleksandar Karamfilov ² and Aleksandar Efremov ¹

¹ Faculty of Automatics, Industrial Automation, Technical University of Sofia, 1000 Sofia, Bulgaria; iveta.grigorova@tu-sofia.bg (I.G.); alefremov@tu-sofia.bg (A.E.)

² Faculty of German Education in Engineering and Business Administration, Technical University of Sofia, 1000 Sofia, Bulgaria; akaramfilov@tu-sofia.bg

* Correspondence: konstantin.piryanov@tu-sofia.bg

Abstract

Ensuring the consistency of recurring ETL processes is a critical challenge in large-scale financial analytics, where upstream data changes—such as variable redefinitions, unit conversions (e.g., from days past due to number of overdue installments or currency changes), or erroneous submissions following source system updates—can silently degrade model reliability. These risks are amplified in automated modeling environments, where dozens of models are retrained monthly for each financial institution and the number of serviced institutions is expected to grow. This study presents an automated statistical monitoring framework for continuous quality assurance of monthly ETL outputs used in model development. The approach quantifies drift between a reference dataset and successive data deliveries using descriptive univariate and bivariate statistics combined with a normalized Canberra-based drift score, aggregated into interpretable variable-level stability measures. Sensitivity is evaluated through controlled noisification experiments with increasing Gaussian perturbations, demonstrating a monotonic decline in stability scores and consistent directional shifts in complementary metrics such as the Gini coefficient and Kolmogorov–Smirnov statistic. The results show that the framework effectively detects both subtle and large-scale distributional changes, providing a scalable, interpretable, and reproducible monitoring diagnostics suitable for fully automated financial data pipelines, with flexibility for extension.

Keywords: data drift monitoring; ETL pipelines; exploratory data analysis; data quality assurance; Canberra distance; statistical monitoring; financial data analytics



Academic Editors: Chengliang Chai and Yu Sun

Received: 20 January 2026

Revised: 14 February 2026

Accepted: 23 February 2026

Published: 26 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Modern data pipelines play a central role in organizational decision-making, supporting reporting systems, analytics platforms, and machine learning models. In financial analytics, these pipelines typically rely on recurring Extract–Transform–Load (ETL) workflows that collect data from multiple external institutions and prepare it for modeling. Such systems operate in dynamic environments: data definitions evolve, source systems change, and business practices vary across institutions. Even small upstream modifications—such as variable redefinitions, unit conversions, or updates in customer databases—can introduce data drift, where statistical properties of the data change over time. If not detected early, these changes may degrade data quality, impair model performance, and undermine trust in automated decision-making systems and weaken governance mechanisms required for trustworthy artificial intelligence and responsible data stewardship [1–4].

In practice, most data quality checks rely on rule-based validation and schema enforcement, including verification of data types, missing values, or allowable ranges. While such checks are essential for identifying structural inconsistencies, they primarily detect violations of predefined constraints and often fail to capture distributional changes that do not explicitly breach validation rules [1–3]. Consequently, statistically meaningful shifts may remain undetected, even though they can significantly affect downstream analytics and predictive models. This limitation is particularly critical in highly automated environments, where manual inspection is infeasible and models are retrained on a recurring basis.

A substantial body of research addresses production-level ML validation, pipeline monitoring, and industrial-scale debugging frameworks [5–8]. In parallel, extensive work on conceptual drift and dataset shift proposes statistical hypothesis tests and distributional distance measures to identify changes in data distributions [9–13]. Although effective in model-centric monitoring scenarios, these methods are often designed to evaluate trained model performance rather than upstream data pipelines. Moreover, they frequently assume fully numeric feature spaces or require computationally intensive distributional procedures. As a result, their direct application to routine ETL monitoring—especially in settings involving heterogeneous feature types and multi-institution data integration—remains challenging in large-scale production systems where validation must scale across diverse pipelines and operational contexts.

Table 1 summarizes the principal strands of related research and highlights a methodological gap. While governance frameworks, model-level drift detection strategies, and theoretical divergence measures have been extensively investigated, comparatively limited attention has been devoted to lightweight, aggregation-based statistical monitoring frameworks specifically tailored to recurring ETL workflows. Existing approaches either emphasize conceptual governance models, focus on model adaptation rather than upstream data validation or rely on full-distribution divergence measures that may be computationally demanding or difficult to interpret in summary-statistic contexts.

To address these limitations, this study proposes an automated statistical monitoring framework explicitly designed for recurring ETL workflows in financial data pipelines. The framework compares each new ETL delivery against a validated reference dataset using a structured collection of univariate and bivariate summary statistics. These metric-level deviations are aggregated through a normalized Canberra-based formulation into bounded, interpretable variable-level stability scores.

We investigate the hypothesis that distributional drift in recurring ETL outputs can be reliably detected using summary statistic-based distance aggregation, without requiring full probability density estimation or access to trained model artifacts. In particular, we examine whether the proposed Canberra-based stability score remains sensitive to both subtle and large-scale distributional perturbations while preserving computational efficiency and interpretability.

Sensitivity is evaluated through controlled noisification experiments in which progressively increasing Gaussian perturbations and categorical perturbation mechanisms are injected into the reference dataset. The results demonstrate consistent and monotonic amplification of stability scores under increasing perturbation intensity, confirming proportional responsiveness to drift magnitude. Complementary validation using real-world ETL cycles further illustrates the framework’s capacity to detect schema evolution, shifts in missingness patterns, and disruptions in inter-variable dependency structures.

Overall, this work demonstrates that statistically grounded, summary-statistic-based aggregation can complement traditional rule-based validation and provide a scalable, interpretable, and operationally practical solution for continuous data quality monitoring in automated financial modeling environments.

Table 1. Summary of related research on data quality, dataset shift, statistical divergence, and production ML monitoring, highlighting methodological focus and limitations in the context of ETL-level statistical drift detection. Source: Authors’ own elaboration.

Category	References	Relative Merits	Limitations in Context of ETL-Level Monitoring
Big Data Quality and Governance	[1–4]	Conceptual foundations of big data quality; systematic detection of data errors; scalable quality verification; governance frameworks for trustworthy AI	Emphasize conceptual models and governance structures; limited focus on automated statistical drift aggregation within recurring ETL processes
ML Pipeline Validation and Debugging	[5–8]	Industrial-scale validation frameworks; production ML debugging; TensorFlow Data Validation; lifecycle-oriented monitoring architectures	Primarily schema-based or rule-driven validation; less emphasis on unified statistical aggregation across heterogeneous summary metrics
Conceptual Drift and Dataset Shift	[9–13]	Formal characterization of drift types; adaptation strategies; monitoring in evolving environments; empirical benchmarking of shift detection methods	Focus mainly on model adaptation or full-distribution shift detection rather than lightweight ETL-level monitoring of scalar summary statistics
Statistical Divergence and Distance Measures	[14–18]	Foundational divergence theory (KL, Jensen-Shannon); optimal transport and Wasserstein distance; classical goodness-of-fit testing (KS)	Often require full probability distributions; may be computationally intensive or less interpretable in heterogeneous ETL summary contexts
High-Dimensional Distance Behavior	[19]	Theoretical analysis of distance concentration and contrast loss in high-dimensional spaces	Not tailored to hierarchical aggregation or ETL operational constraints
MLOps and Production Monitoring Architectures	[20,21]	Drift mitigation in real-time systems; taxonomy of modular MLOps architectures	Limited integration of unified statistical monitoring layers specifically designed for ETL-generated summary metrics.

2. Materials and Methods

2.1. ETL Context and System Architecture

The proposed monitoring framework operates as a post-processing component within an existing Extract–Transform–Load (ETL) pipeline used for financial data modeling. Its primary role is to verify the statistical consistency of recurring ETL outputs against a predefined reference dataset, representing validated, business-approved behavior. The framework is positioned immediately after exploratory data analysis (EDA) and prior to model training, serving as an automated quality gate for downstream analytics.

The system architecture, illustrated in Figure 1, is intentionally decoupled into two operational layers:

- **Metric Extraction (ETL Layer)**—During routine ETL execution, summary statistics are computed for both the reference dataset D_{ref} and the incoming dataset D_t . This stage performs the data-intensive operations, including distributional profiling and dependency estimation, and scales linearly with dataset size.
- **Drift Scoring (Monitoring Layer)**—The monitoring engine operates exclusively on the stored summary statistics rather than raw observations. By comparing reference and current metric vectors using the Canberra-based aggregation procedure, drift detection becomes asymptotically independent of dataset cardinality. Monitoring latency therefore depends only on the number of variables and metrics, rather than on the number of records.

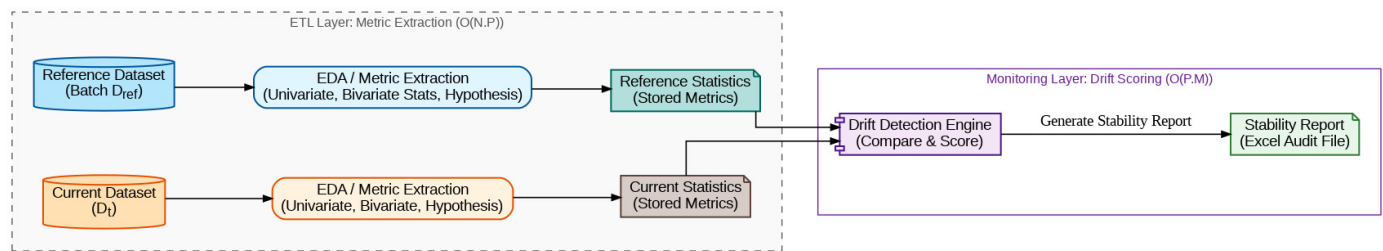


Figure 1. Architecture of the proposed ETL drift monitoring framework. The system separates the data-intensive metric extraction phase (ETL Layer) from the lightweight drift scoring phase (monitoring Layer). Summary statistics computed for the reference dataset D_{ref} and the current dataset D_t are compared using normalized Canberra-based aggregation to produce variable-level stability scores. Source: Authors' own elaboration.

ETL outputs may be provided as comma-separated value (CSV) files or as serialized Python objects (pickle files) containing precomputed EDA metrics. When summary statistics are already available, the monitoring layer bypasses raw-data processing entirely, further reducing computational overhead.

The framework employs a fixed reference window strategy. Each incoming ETL delivery D_t is compared against a static, validated reference dataset D_{ref} , ensuring that drift is assessed relative to a stable baseline rather than through comparisons between consecutive ETL runs. This design mitigates gradual drift masking, whereby incremental distributional changes might otherwise be absorbed into a continuously updated baseline.

The reference dataset is reviewed and updated following major ETL releases, schema revisions, or substantial upstream logic changes. This controlled update policy preserves sensitivity to cumulative structural and distributional drift while maintaining operational realism. The fixed reference window strategy is particularly suited to batch-oriented ETL workflows, such as monthly financial deliveries, where complete dataset snapshots are evaluated at discrete intervals. Exploration of rolling or adaptive reference strategies is left for future work.

2.2. Input Handling and Configuration

System behavior is governed by a centralized configuration workbook specifying input paths, variable definitions, semantic data types, optional preprocessing rules, and modeling strategies. The configuration notebook enables consistent schema interpretation across ETL cycles, supports manual overrides (e.g., reclassifying continuous variables as ordinal), and provides traceability for auditing and reproducibility.

When CSV inputs are provided, the system executes an ETL run and computes the required statistical metrics. When serialized EDA outputs are available, metrics are loaded directly to enable efficient reuse of prior computations. All processed outputs are stored as standardized, serialized dictionaries to ensure consistency and low input–output latency.

2.3. Schema Alignment and Metric Extraction

Prior to statistical comparison, schema alignment is performed between the reference dataset and the current ETL output. Added, removed, or missing variables are explicitly identified and reported. Only features present in both datasets are included in drift aggregation.

For aligned variables, the framework constructs metric tables across three categories:

- Univariate metrics—number of unique values, missing count and proportion, mean, median, standard deviation, and quartiles.
- Bivariate metrics—Pearson, Spearman, and Kendall correlation coefficients; Gini index, Kolmogorov–Smirnov statistic; Diversity index; and trend indicators [13,16,18].

- Hypothesis-based metrics— p -values from chi-square tests and correlation significance tests.

Standard statistical metrics are computed using established implementations in widely used Python libraries.

2.4. Metric Definitions

All metrics are computed at the variable level and form the basis for subsequent aggregation and scoring.

2.4.1. Absolute and Relative Deviations

Let x^r and x^c denote the value of a given metric computed on the reference and the current ETL datasets, respectively.

The absolute deviation is defined as:

$$\Delta_{abs} = |x^r - x^c|. \quad (1)$$

The relative deviation is defined as:

$$\Delta_{rel} = \frac{x^c - x^r}{|x^r|} \times 100, \quad (2)$$

When $x^r = 0$, the relative deviation may become numerically unstable. Such cases are retained in diagnostic outputs but excluded from normalized aggregation steps requiring finite values. Relative deviations are used for interpretive diagnostics and are not directly aggregated into the Canberra-based stability score.

2.4.2. Overlap-Based Diversity Index

Let $g_i \geq 0$ and $b_i \geq 0$ denote the counts (or non-negative weights) of two groups across units $i = 1, \dots, n$, with total masses

$$G = \sum_{i=1}^n g_i, \quad B = \sum_{i=1}^n b_i. \quad (3)$$

The overlap-based diversity (dissimilarity) index is defined as

$$D = 100 \left(1 - \sum_{i=1}^n \frac{g_i b_i}{g_i + b_i} \times \frac{G + B}{GB} \right) \quad (4)$$

Units for which $g_i + b_i = 0$ contribute zero to the sum; in implementation, a small numerical constant is used to ensure stability without affecting the limit behavior.

The index takes values in the interval $[0, 100]$, where lower values indicate stronger co-location (overlap) of the two groups across units, and higher values indicate greater separation. The normalization factor $(G + B)/(GB)$ ensures scale invariance and comparability across datasets with different marginal totals.

Unlike classical diversity indices based solely on marginal proportions, this measure is sensitive to joint distributional structure and local co-occurrence patterns through harmonic-type weighting. The index is symmetric in the two groups, invariant under uniform scaling of counts, bounded, and continuous with respect to small perturbations of the data, making it suitable for monitoring distributional change over time.

2.4.3. Expected Trend Indicator

For selected metrics, the framework incorporates an expected trend indicator τ , specified externally via the configuration file. The indicator encodes the anticipated directional

behavior of a metric relative to the reference dataset and reflects domain knowledge rather than empirical estimation.

$$\tau \in \{-1, 0, 1\}, \quad (5)$$

where:

- $\tau = 1$ indicates an expected increase,
- $\tau = -1$ indicates an expected decrease,
- $\tau = 0$ indicates no expected directional change.

The indicator does not modify the Canberra distance formulation or the magnitude of component-wise deviations. Instead, it functions as a contextual attribute within the bivariate monitoring component, enabling differentiation between anticipated directional evolution and potentially anomalous change during interpretation and reporting.

It therefore supports semantically informed drift diagnostics without introducing thresholding rules, hypothesis constraints, or additional weighting terms into the aggregation procedure.

2.4.4. Standard Bivariate Metrics

Pearson, Spearman, and Kendall correlation coefficients, as well as the Gini coefficient and Kolmogorov–Smirnov statistic [18], are computed using their standard definitions as implemented in widely used Python statistical libraries. These metrics complement the custom measures by capturing linear, monotonic, ordinal, and distributional relationships between variables.

2.5. Data Preprocessing and Variable Treatment

The monitoring framework applies type-aware preprocessing to ensure numerically stable and statistically consistent metric extraction across ETL cycles. Because the objective is distributional monitoring rather than predictive modeling, preprocessing is deliberately minimal and avoids transformations that would obscure raw statistical structure.

2.5.1. Feature Scaling

No global feature scaling (e.g., z-score normalization or min-max transformation) is applied to raw variables prior to metric computation. All univariate statistics (mean, median, standard deviation, quartiles) are calculated in the native units of each feature to preserve interpretability. Cross-metric comparability is achieved at the aggregation stage through the normalized Canberra formulation (Section 2.6), which ensures that heterogeneous metrics contribute proportionately regardless of scale.

2.5.2. Handling of Missing and Non-Finite Values

Missingness is treated as an explicit monitoring signal rather than a nuisance variable. For each feature, the framework computes:

- The total number of missing observations
- The proportion of missing values
- The number of unique values including missing entries

Continuous summary statistics are computed only on valid, finite observations. Specifically, values equal to NaN, $+\infty$, or $-\infty$ are excluded from mean, median, standard deviation, and quantile calculations. If no valid values remain after filtering, the corresponding statistics are recorded as NaN to preserve diagnostic transparency. These undefined values are retained in the audit output but excluded from normalized aggregation where finite inputs are required.

2.5.3. Treatment of Numerical, Categorical, and Ordinal Variables

The framework supports heterogeneous feature types without requiring one-hot encoding or embedding transformations.

For numerical variables, summary statistics are computed directly on raw values. For categorical variables, value distributions are derived using full frequency tables (including missing categories), ensuring that changes in level prevalence or missingness are explicitly captured. Ordinal variables are treated analogously but retain their inherent ordering during correlation computations.

When computing correlation coefficients (Pearson, Spearman, Kendall) against the configured dependent variable, categorical features are temporarily mapped to integer category codes to enable numeric correlation computation. This mapping is used solely for statistical calculation and does not modify the stored variable representation.

2.5.4. High-Cardinality Features and Binning

Variables exceeding a configurable unique-value threshold are flagged for binning to stabilize distributional comparisons and mitigate sparsity effects. Binning is performed through the configured binning module using minimum-record constraints per bin and a fixed random seed for reproducibility.

For binned variables, value-distribution summaries and downstream dependency metrics (Gini index, KS statistic, Diversity index and Chi-square statistics) are computed at the bin level. Missing values are explicitly assigned to a dedicated bin to preserve interpretability and maintain consistency across ETL runs.

All reported bivariate statistics are rounded according to the configurable precision parameter. Numerical stabilization is applied where required to prevent division-by-zero artifacts while preserving asymptotic behavior.

2.6. Canberra-Based Scoring Framework

To summarize metric-level deviations into interpretable drift indicators, the framework employs a normalized Canberra-style distance [14,15]. The objective is to aggregate heterogeneous univariate and bivariate monitoring metrics into a single, comparable drift score at the variable level.

Let v denote a variable. The set of monitoring metrics associated with v is divided into two subsets: univariate metrics (M_{uni}) and bivariate metrics (M_{biv}). For each metric m , let x^r and x^c denote the calculated metric values derived from the reference (benchmark) and current ETL datasets respectively.

2.6.1. Metric Normalization and Scaling

Prior to aggregation, metric values are normalized to ensure numerical stability and comparability across metrics with heterogeneous scales.

For the univariate metric CountUV (number of unique values), reference and current values are normalized using their joint mass. Let x^r and x^c represent the raw reference and current counts. The normalized values $\hat{x}$ are defined as:

$$\hat{x}^r = \frac{x^r}{x^r + x^c + \epsilon}, \quad \hat{x}^c = \frac{x^c}{x^r + x^c + \epsilon} \quad (6)$$

where $\epsilon = 10^{-12}$ is a small constant introduced to ensure numerical stability.

If $x^r = x^c = 0$, both normalized values are set to 0.5, yielding a zero Canberra contribution and preventing spurious drift.

For selected bivariate metrics that are stored on a percentage scale—specifically the Gini coefficient, Diversity index, and Kolmogorov–Smirnov statistic—values are rescaled

by dividing by 100 prior to aggregation (i.e., $\hat{x} = x / 100$, where x is the raw percentage). All remaining metrics are used in their native numeric form. All normalization and scaling operations are applied symmetrically to reference and current values.

2.6.2. Canberra-Based Drift Aggregation

Given the normalized metric values $\hat{x}^r$ and $\hat{x}^c$, the Component-wise Canberra Distance d_m for variable v on metric m is defined as:

$$d_m(\hat{x}^r, \hat{x}^c) = \frac{|\hat{x}^r - \hat{x}^c|}{|\hat{x}^r| + |\hat{x}^c| + \epsilon} \quad (7)$$

Each term in the summation is bounded in the interval $[0, 1]$. To prevent high-dimensional univariate statistics from diluting signals of relationship drift, sub-scores are computed separately for univariate and bivariate dimensions.

$$S_{uni} = \frac{1}{|M_{uni}|} \sum_{m \in M_{uni}} d_m, \quad S_{biv} = \frac{1}{|M_{biv}|} \sum_{m \in M_{biv}} d_m \quad (8)$$

The final Variable Stability Score S_v is computed as the arithmetic mean of these two components:

$$S_v = \frac{S_{uni} + S_{biv}}{2} \quad (9)$$

This hierarchical aggregation ensures that structural dependencies (captured by bivariate metrics) and statistical properties (captured by univariate metrics) contribute equally to the final drift assessment. The overall scoring workflow is illustrated in Figure 2.

In the current formulation, all variables are treated with equal importance when computing drift scores, irrespective of business relevance, downstream predictive importance, or historical volatility. This design choice is intentional and reflects the framework's role as a generic, model-agnostic monitoring layer within the ETL pipeline. By avoiding variable-specific weighting, the framework ensures consistent, transparent drift detection across heterogeneous datasets without requiring access to proprietary business knowledge or trained downstream models.

Importantly, equal weighting prevents high-variance or highly predictive variables from dominating drift assessments, thereby preserving sensitivity to distributional changes in less prominent features that may nonetheless signal upstream data quality issues. Incorporating business- or model-driven weights is a natural extension of the framework and can be achieved by modifying the final aggregation step; however, such strategies are deliberately deferred to future work to preserve interpretability and broad applicability in diverse operational settings.

2.6.3. Rationale for Canberra-Based Formulation

The Canberra-based formulation was selected due to its suitability for aggregating heterogeneous summary statistics commonly produced in ETL monitoring workflows. Unlike divergence measures such as Jensen-Shannon or Wasserstein distance, the Canberra distance operates directly on scalar descriptors and emphasizes relative deviations normalized by the magnitude of the compared values. This property is particularly advantageous when monitoring metrics with different units, scales, and baseline variability, including cases where benchmark values approach zero.

Distance-based measures are known to exhibit altered behavior in high-dimensional settings, where distance concentration and reduced contrast between observations may limit discriminative power [19]. In the context of ETL monitoring, where drift assessment involves aggregating multiple heterogeneous metrics across potentially hundreds of variables, such

effects could diminish sensitivity if distances are applied directly to raw feature vectors. The proposed formulation mitigates this risk by operating on normalized, interpretable summary statistics and by separating univariate and bivariate components prior to aggregation. This hierarchical structure preserves metric-level contrast while avoiding the curse-of-dimensionality effects associated with direct high-dimensional distance computation.

In addition, the Canberra distance is computationally lightweight, interpretable at the individual metric level, and naturally extensible to variable-level aggregation, making it well aligned with the operational constraints of automated, large-scale ETL pipelines.

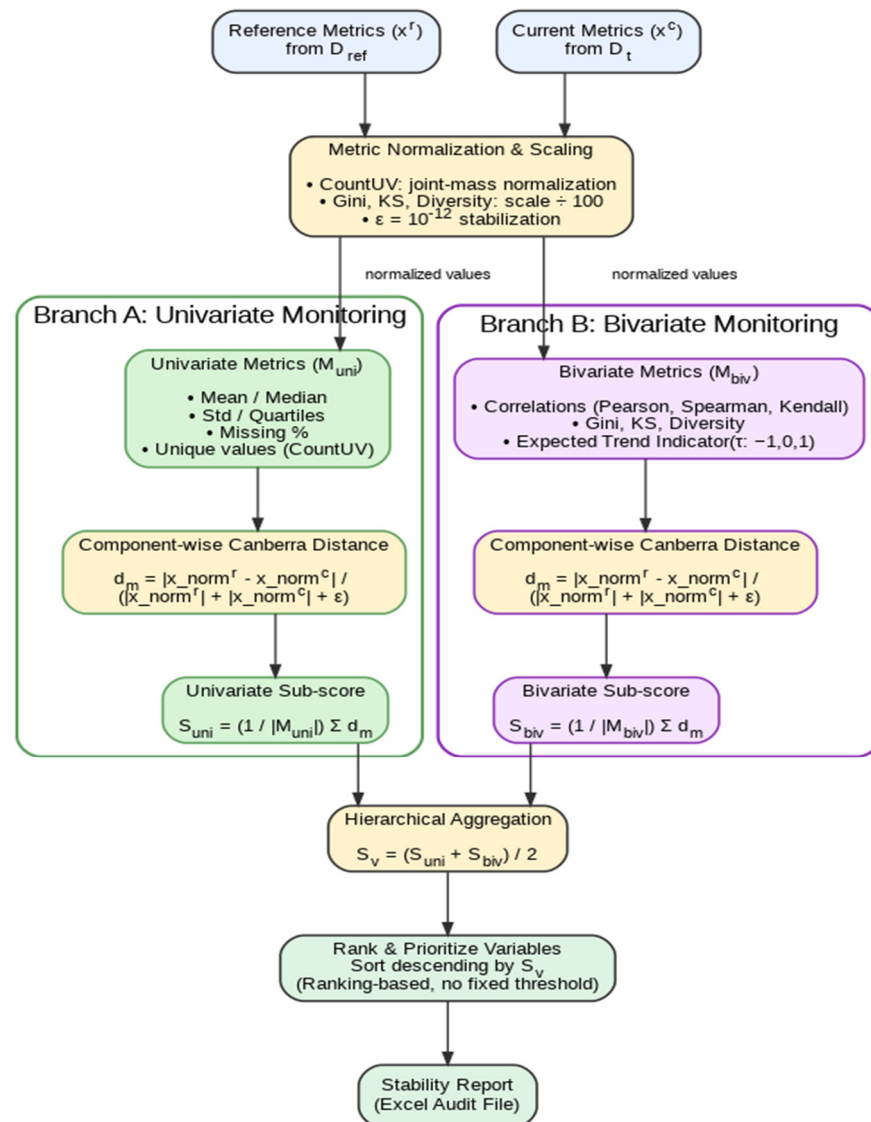


Figure 2. Hierarchical Canberra-based drift scoring mechanism. The framework processes reference (x^r) and current (x^c) metrics through normalization (Equation (6)) before splitting them into univariate and bivariate dimensions. Component-wise Canberra distances are computed (Equation (7)) and averaged into dimension-specific sub-scores (Equation (8)). These are finally aggregated via arithmetic mean (Equation (9)) into a global variable stability score (S_v), which is used to rank variables by drift severity. Source: Authors' own elaboration.

2.6.4. Alerting and Thresholding Strategy

Unlike binary classifiers that require fixed decision boundaries, the proposed framework operates as a ranking-based monitoring system. This approach mitigates alert fatigue, a common failure mode in high-dimensional financial datasets where static thresholds produce excessive false positives.

While no universal threshold is imposed a priori, an initial heuristic baseline can be derived from sensitivity analysis: the experimental results in Section 3.5. indicate that stability scores exceeding approximately 0.20 consistently correspond to meaningful distributional changes under medium-intensity perturbations. This value therefore serves as a starting reference point for prioritization rather than a strict alarm boundary.

Threshold interpretation is adapted over time based on historical score distributions, dataset characteristics, and operational feedback. As the monitoring system accumulates longitudinal evidence across ETL cycles, score percentiles, trends, or stability score dispersion can be used to recalibrate prioritization criteria. This adaptive interpretation allows expected data evolution to be distinguished from anomalous drift without enforcing rigid, static thresholds.

2.7. Synthetic Noisification for Sensitivity Analysis

To evaluate the sensitivity of the proposed monitoring framework under controlled drift [9–13], synthetic perturbations were applied to the reference (benchmark) dataset using a reproducible noisification module with a fixed random seed ($s = 42$). The experiments focus on variance-based perturbations for continuous variables and distributional perturbations for categorical variables, which represent common and general forms of distributional drift in recurring ETL workflows.

Three noise intensity levels—low, medium and high—were defined via a variance control parameter. The parameters used for each noise level are summarized in Table 2.

$$p_{var} \in \{0.1, 0.3, 0.5\}. \quad (10)$$

Table 2. Parametrization of synthetic noisification used in sensitivity analysis. Noise intensity is controlled by perturbation probability p_{var} , with type-specific perturbation mechanisms applied to continuous (Gaussian noise), binary/nominal (random flipping), and ordinal variables (discrete shift probabilities). Source: Authors’ own elaboration.

Noise Level	p_{var}	Continuous Perturbation	Binary/Nominal Flip Probability	Ordinal Shift Probabilities
Low	0.1	$\sigma_x = s_x \sqrt{0.1}$	0.10	$P(\pm 1) = 0.05, P(0) = 0.90$
Medium	0.3	$\sigma_x = s_x \sqrt{0.3}$	0.30	$P(\pm 1) = 0.15, P(0) = 0.70$
High	0.5	$\sigma_x = s_x \sqrt{0.5}$	0.50	$P(\pm 1) = 0.25, P(0) = 0.50$

For continuous variables, variance inflation was simulated by exclusively adding zero-mean Gaussian noise,

$$x' = x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_x^2), \quad (11)$$

where the feature-specific standard deviation is defined as

$$\sigma_x = s_x \sqrt{p_{var}}, \quad (12)$$

and s_x denotes the sample standard deviation of the variable in the benchmark dataset. This feature-adaptive formulation ensures that injected noise scales proportionally with baseline variability across heterogeneous features.

For categorical variables, distributional drift was simulated through random value perturbations applied to a fraction p_{var} of non-missing observations. Binary variables were flipped to the opposite category with probability p_{var} . Nominal variables were reassigned to a different observed category selected uniformly from the remaining levels. Ordinal variables were shifted by ± 1 with probabilities $p_{var} / 2$ each (and left unchanged otherwise), with resulting values clipped to the observed range.

Each perturbed dataset was processed through the same exploratory data analysis and monitoring pipeline as the unperturbed reference dataset. Drift scores were then computed relative to the reference dataset, enabling assessment of score behavior as a function of controlled perturbation intensity.

This experimental design isolates sensitivity to fundamental distributional changes in both numerical and categorical features, providing a focused and interpretable validation of the framework's responsiveness to increasing levels of synthetic drift.

The synthetic experiments operate exclusively on perturbed versions of the benchmark dataset D_{ref} ; the current ETL dataset D_t is used only for real-world validation and is not involved in the noisification procedure.

Dataset Characteristics and Implementation Details

The reference dataset D_{ref} contained 99,257 observations across 415 variables. The current ETL dataset D_t consisted of 251,211 observations and 412 variables. Prior to stability score computation, schema alignment was performed to identify added, removed, or missing variables. Only features present in both datasets were included in drift aggregation. The dimensionality difference reflects natural schema evolution between ETL cycles.

The monitored variables comprised a heterogeneous mix of numerical, categorical, and ordinal features typical of financial data modeling workflows. The reference dataset preceded the evaluation dataset chronologically and was treated as the validated reference state under the fixed-window monitoring strategy described in Section 2.1.

The framework was implemented in Python (version 3.9) using pandas (version 1.4.3) for data manipulation, NumPy (version 1.25.0) for numerical operations, and SciPy (version 1.10.1) for statistical computations. Pearson, Spearman, and Kendall correlations were computed using standard SciPy and pandas implementations. The two-sample KS statistic was calculated using SciPy's two-sided test. All synthetic perturbations were generated using NumPy's random number generator with fixed seed 42. The Canberra normalization stabilization constant was fixed at $\epsilon = 10^{-12}$ throughout all experiments. Unless otherwise specified, full-batch datasets were processed without subsampling.

2.8. Evaluation Criteria and Performance Metrics

The synthetic perturbation experiments described in Section 2.6 provide a controlled setting for evaluating the framework's drift detection behavior; the following section defines the criteria used to assess performance under these conditions.

Because the proposed framework produces continuous, diagnostic drift scores rather than binary drift decisions, performance evaluation differs from that of supervised drift detectors. Traditional detection metrics such as accuracy, precision, recall, or detection delay are not directly applicable in the absence of labeled drift events or a predefined alarm threshold. Instead, evaluation focuses on the following criteria, which reflect the intended role of the framework as an ETL monitoring and diagnostic tool:

- **Sensitivity to Distributional Change**—performance is assessed by the framework's ability to produce increasing drift scores in response to progressively stronger distributional perturbations. This property is evaluated through controlled noisification experiments, where monotonic increases in the Canberra-based stability score indicate effective sensitivity to drift magnitude.
- **Robustness Across Data Types and Metrics**—the framework's performance is further assessed by its consistent behavior across numerical, categorical, and ordinal variables, as well as across univariate and bivariate metrics. Stable variables are expected to exhibit low scores, while structurally altered variables are ranked higher regardless of data type.

- Discriminative Ranking ability—rather than triggering binary alarms, the framework prioritizes variables by drift severity. Performance is therefore evaluated by the clarity and consistency of this ranking, enabling practitioners to focus attention on the most affected variables during investigation.
- Interpretability and Diagnostic Value—an additional performance criterion is interpretability: drift scores must be traceable to underlying metric deviations and understandable to practitioners. The use of bounded, normalized scores and audit tables supports transparent root-cause analysis, which is critical in operational ETL monitoring contexts.

Together, these criteria provide a comprehensive and task-appropriate evaluation of drift detection performance in unsupervised, production-oriented monitoring settings.

2.9. Reporting and Outputs

Results are consolidated into a standardized Excel report designed for both automated review and human inspection. The report includes the following sheets:

- Benchmark metrics
- Current-run metrics
- Absolute Deviations
- Canberra audit table (interleaved benchmark and current values used for calculating it)
- Variable-level and aggregate drift scores
- Missing and schema-mismatch summaries

Consistent formatting, column structure, and highlighting are applied to improve readability and support operational use.

2.10. Validation

The framework was evaluated using both synthetic and real-world data. Controlled noisification experiments demonstrated a monotonic increase in Canberra-based scores with increasing noise intensity, confirming sensitivity to distributional changes. Historical ETL case studies further showed that the system reliably identified schema changes, shifts in missingness patterns, and altered dependency structures. These results validate the framework's effectiveness as an early warning system for both structural and statistical drift in recurring ETL workflows.

3. Results

Given the large number of variables produced by the ETL pipeline, results are reported for a representative subset of features to preserve clarity and readability. At each analysis stage, we present the top five and bottom five variables ranked by the final Canberra-based stability score. This selection highlights features exhibiting the strongest and weakest evidence of drift, respectively.

Results are organized in four stages that reflect the workflow of the framework: baseline characterization of benchmark and current datasets, deviations at the metric level (absolute and relative), aggregated Canberra-based stability scores, and controlled noisification experiments to assess sensitivity to synthetic drift. We additionally report an ablation analysis of monitoring components and two anonymized operational case studies.

3.1. Baseline Dataset Characterization

Tables 3 and 4 summarize univariate descriptive statistics for the reference and current datasets, focusing on the top and bottom five variables ranked by final Canberra-based score.

Table 3. Univariate descriptive statistics for the top five and bottom five variables ranked by final Canberra-based stability score in the reference dataset D_{ref} . Source: Authors' own elaboration.

Variable	Unique Values	Missing Count	Mean	Median	STD	p25	p50	p75
Top 1	4	0	0.0001	0	0.013	0	0	0
Top 2	4	0	0.0001	0	0.0114	0	0	0
Top 3	4	0	0.0028	0	0.3719	0	0	0
Top 4	42	0	0.6904	0	50.7769	0	0	0
Top 5	15	0	0.005	0	0.2184	0	0	0
Bottom 1	1	0	1	1	0	1	1	1
Bottom 2	2	0	-	-	-	-	-	-
Bottom 3	2	0	-	-	-	-	-	-
Bottom 4	2	0	-	-	-	-	-	-
Bottom 5	2	0	-	-	-	-	-	-

Table 4. Univariate descriptive statistics for the top five and bottom five variables ranked by final Canberra-based stability score in the current dataset D_t . Source: Authors' own elaboration.

Variable	Unique Values	Missing Count	Mean	Median	STD	p25	p50	p75
Top 1	1663	0	1.5471	0.77	2.7287	0.36	0.77	1.7
Top 2	1443	0	1.2885	0.68	2.1537	0.33	0.68	1.45
Top 3	33	0	17.149	15	19.4206	15	15	15
Top 4	109	0	11.5068	0	519.7718	0	0	0
Top 5	18	0	0.01	0	0.2958	0	0	0
Bottom 1	1	0	1	1	1	1	1	1
Bottom 2	2	0	-	-	-	-	-	-
Bottom 3	2	0	-	-	-	-	-	-
Bottom 4	2	0	-	-	-	-	-	-
Bottom 5	2	0	1	1	0	1	1	1

In the reference dataset (Table 3), the top-ranked variables exhibit low baseline variability, narrow interquartile ranges, and negligible missingness. Such variables are expected to be sensitive to changes in ETL logic because even moderate distributional shifts can produce measurable deviations in summary statistics. In contrast, the bottom-ranked variables are predominantly categorical with one or two unique values; for these features, numerical summary statistics are constant or undefined, consistent with identifier-like behavior or binary flags exhibiting minimal temporal variability across ETL cycles.

In the current dataset summarized in Table 4, the overall schema is largely preserved. However, the top-ranked variables exhibit increases in the number of unique values and greater dispersion in central tendency and spread measures. These changes indicate distributional broadening relative to the reference dataset and provide an early signal of drift. Conversely, the bottom-ranked categorical variables retain nearly identical distributions, confirming schema integrity and temporal stability.

3.2. Bivariate Structure Analysis

Tables 5 and 6 summarize bivariate metrics for the reference and current datasets. In reference data (Table 5), top-ranked variables show weak dependencies with respect to the configured dependent variable, with correlation coefficients close to zero and low inequality and divergence measures. This indicates limited inter-variable structure at the reference state. The bottom-ranked variables exhibit stronger and more stable dependency patterns, reflected in higher correlation and higher Gini, KS and Diversity values, which is typical for categorical or ordinal features with dominant levels or fixed relationships.

Table 5. Bivariate monitoring metrics (correlation, inequality/divergence, diversity, and trend) for the top five and bottom five variables ranked by final Canberra-based stability score in the reference dataset D_{ref} . Source: Authors' own elaboration.

Variable	Pearson Cor.	Spearman Cor.	Kendall Cor.	Gini	KS	Diversity	Trend
Top 1	−0.0113	−0.0137	−0.0137	0	0	0	-
Top 2	−0.0114	−0.0137	−0.0137	0	0	0	-
Top 3	−0.0137	−0.0137	−0.0137	0	0	0	-
Top 4	0.0041	0.0023	0.0023	0	0	0	-
Top 5	0.0053	0.0023	0.0023	0	0	0	-
Bottom 1	-	-	-	0	0	0	-
Bottom 2	0.5702	0.5702	0.5702	63.207	63.207	32.5113	-
Bottom 3	0.4527	0.4527	0.4527	52.7265	52.7265	20.4913	-
Bottom 4	0.516	0.516	0.516	59.0617	59.0617	26.6305	-
Bottom 5	−0.0327	−0.0327	−0.0327	3.6653	3.6653	0.1067	-

Table 6. Bivariate monitoring metrics for the top five and bottom five variables ranked by final Canberra-based stability score in the current dataset D_t . Source: Authors' own elaboration.

Variable	Pearson Cor.	Spearman Cor.	Kendall Cor.	Gini	KS	Diversity	Trend
Top 1	0.006	0.0247	0.0204	3.5324	3.9118	0.6505	-
Top 2	0.0056	0.0246	0.0203	3.5117	3.9707	0.584	-
Top 3	−0.0234	−0.0069	−0.0067	0.6418	1.8401	0.26	-
Top 4	−0.0401	−0.045	−0.045	0.473	0.473	0.202	-
Top 5	−0.0262	−0.0449	−0.0449	0.473	0.473	0.202	-
Bottom 1	-	-	-	0	0	0	-
Bottom 2	0.5673	0.5673	0.5673	63.1409	63.1409	32.1864	-
Bottom 3	0.4517	0.4517	0.4517	53.536	53.536	20.4056	-
Bottom 4	0.5176	0.5176	0.5176	59.9667	59.9667	26.7877	-
Bottom 5	−0.0327	−0.0327	−0.0327	3.7549	3.7549	0.1069	-

In the current dataset summarized in Table 6, bivariate relationships among the bottom-ranked variables remain largely unchanged. For top-ranked variables, small but consistent shifts appear across correlation and inequality/divergence measures, suggesting changes in dependency structure and/or distributional shape relative to the reference dataset.

3.3. Deviation Analysis: Absolute and Relative Deviations

Table 7 reports absolute univariate differences between the reference and current datasets. Variables with the highest Canberra scores exhibit pronounced changes in key statistics such as mean, standard deviation, and quartiles, consistent with increased variability in the current ETL output. Bottom-ranked variables show minimal or zero absolute differences.

Table 8 reports absolute bivariate differences. Top-ranked variables display larger deviations across correlation and inequality/divergence measures, indicating changes in dependency structure. Deviations among bottom-ranked variables remain small.

Relative differences are reported in Tables 9 and 10. For top-ranked variables, relative deviations can be very large when the reference metric is near zero, yielding inflated ratios. These values are retained for diagnostic interpretation because they highlight metrics with unstable or degenerate baselines; however, the final drift score is computed using the normalized Canberra-based formulation described in Section 2.6, which includes stabilization (ϵ) and bounded aggregation. Bottom-ranked variables exhibit negligible relative differences. Undefined values arise from zero denominators and follow standard numerical behavior (e.g., NumPy propagation), and are reported accordingly.

Table 7. Absolute univariate deviations Δ_{abs} between the reference dataset D_{ref} and the current dataset D_t for the top five and bottom five variables ranked by the final Canberra-based stability score. Source: Authors' own elaboration.

Variable	Unique Values	Missing Count	Mean	Median	STD	p25	p50	p75
Top 1	1659	0	1.547	0.77	2.7157	0.36	0.77	1.7
Top 2	1439	0	1.2885	0.68	2.1423	0.33	0.68	1.45
Top 3	29	0	17.1462	15	19.0487	15	15	15
Top 4	67	0	10.8164	0	468.9949	0	0	0
Top 5	3	0	0.0051	0	0.0774	0	0	0
Bottom 1	0	0	0	0	0	0	0	0
Bottom 2	0	0	-	-	-	-	-	-
Bottom 3	0	0	-	-	-	-	-	-
Bottom 4	0	0	-	-	-	-	-	-
Bottom 5	0	0	-	-	-	-	-	-

Table 8. Absolute bivariate deviations Δ_{abs} between the reference dataset D_{ref} and the current dataset D_t for the top five and bottom five variables ranked by the final Canberra-based stability score. Source: Authors' own elaboration.

Variable	Pearson Cor.	Spearman Cor.	Kendall Cor.	Gini	KS	Diversity	Trend
Top 1	0.0173	0.0384	0.0341	3.5324	3.9118	0.6505	-
Top 2	0.017	0.0383	0.034	3.5117	3.9707	0.584	-
Top 3	0.0097	0.0068	0.007	0.6418	1.8401	0.26	-
Top 4	0.0442	0.0473	0.0473	0.473	0.473	0.202	-
Top 5	0.0315	0.0472	0.0472	0.473	0.473	0.202	-
Bottom 1	-	-	-	0	0	0	-
Bottom 2	0.0029	0.0029	0.0029	0.0661	0.0661	0.3249	-
Bottom 3	0.001	0.001	0.001	0.8095	0.8095	0.0857	-
Bottom 4	0.0016	0.0016	0.0016	0.905	0.905	0.1572	-
Bottom 5	0	0	0	0.0896	0.0896	0.0002	-

Table 9. Relative univariate deviations Δ_{rel} (%) between the reference dataset D_{ref} and the current dataset D_t for the top five and bottom five variables ranked by the final Canberra-based stability score. Source: Authors' own elaboration.

Variable	Unique Values	Missing Count	Mean	Median	STD	p25	p50	p75
Top 1	41,475	-	1,933,718.84	-	20,893.76	-	-	-
Top 2	35,975	-	1,809,179.074	-	18,754.106	-	-	-
Top 3	725	-	615,587.8004	-	5121.3275	-	-	-
Top 4	159.5238	-	1566.6069	-	923.6389	-	-	-
Top 5	20	-	102.142	-	35.4542	-	-	-
Bottom 1	0	-	0	0	-	0	0	0
Bottom 2	0	-	-	-	-	-	-	-
Bottom 3	0	-	-	-	-	-	-	-
Bottom 4	0	-	-	-	-	-	-	-
Bottom 5	0	-	-	-	-	-	-	-

Table 10. Relative bivariate deviations Δ_{rel} (%) between the reference dataset D_{ref} and the current dataset D_t for the top five and bottom five variables ranked by the final Canberra-based stability score. Source: Authors' own elaboration.

Variable	Pearson Cor.	Spearman Cor.	Kendall Cor.	Gini	KS	Diversity	Trend
Top 1	153.0973	280.292	248.9051	-	-	-	-
Top 2	149.1228	279.562	248.1752	-	-	-	-

Table 10. Cont.

Variable	Pearson Cor.	Spearman Cor.	Kendall Cor.	Gini	KS	Diversity	Trend
Top 3	70.8029	49.635	51.0949	-	-	-	-
Top 4	1078.0488	2056.5217	2056.5217	-	-	-	-
Top 5	594.3396	2052.1739	2052.1739	-	-	-	-
Bottom 1	-	-	-	-	-	-	-
Bottom 2	0.5086	0.5086	0.5086	0.1046	0.1046	0.9993	-
Bottom 3	0.2209	0.2209	0.2209	1.5353	1.5353	0.4182	-
Bottom 4	0.3101	0.3101	0.3101	1.5323	1.5323	0.5903	-
Bottom 5	0	0	0	2.4445	2.4445	0.1874	-

3.4. Canberra-Based Stability Scores

Table 11 reports the aggregated univariate score (S_{uni}), bivariate score (S_{biv}), and final stability score (S_v). Top-ranked variables show substantially higher final scores, reflecting combined evidence from univariate dispersion and changes in bivariate structure. Bottom-ranked variables achieve near-zero scores, indicating negligible drift.

Table 11. Aggregated univariate score S_{uni} , bivariate score S_{biv} , and final stability score S_v for the top five and bottom five variables ranked by the final Canberra-based stability score. Source: Authors' own elaboration.

Variable	Univariate Score	Bivariate Score	Final Score
Top 1	0.8551	0.8571	0.8561
Top 2	0.8548	0.8571	0.856
Top 3	0.8208	0.5621	0.6915
Top 4	0.3075	0.8571	0.5823
Top 5	0.0828	0.8571	0.47
Bottom 1	0	0	0
Bottom 2	0	0.002	0.001
Bottom 3	0	0.0029	0.0015
Bottom 4	0	0.0033	0.0016
Bottom 5	0	0.0036	0.0018

These results confirm that the Canberra-based aggregation yields an interpretable ranking of variables by drift severity and separates stable features from those undergoing meaningful distributional or relational change.

3.5. Experimental Noisification Results

To assess sensitivity under controlled conditions, synthetic perturbations were applied to the reference dataset at low, medium, and high intensity levels (Section 2.7). Each perturbed dataset was compared directly to the reference using the same metric extraction and Canberra-based aggregation procedure.

Table 12 reports results for the medium-intensity condition, focusing on Gini and KS deviations and the resulting stability scores. Noise injection increases inequality and distributional divergence metrics, particularly for variables with high baseline stability, yielding final stability scores approaching 0.8. This indicates that the framework responds proportionally to controlled perturbations.

To assess score behavior across perturbation levels, Figures 3–5 provide complementary visualizations. Figure 3 shows the distribution of final stability scores under low, medium and high perturbation intensity. Increasing perturbation strength shifts the distribution upward and increases dispersion. Figure 4 reports the median stability score as a function of intensity and shows a monotonic increase across levels. Figure 5 illus-

trates variable-level stability trajectories across perturbation levels, indicating the upward response is distributed across many features rather than driven by a small subset.

Table 12. Medium-intensity synthetic noisification ($p_{var} = 0.3$): absolute and relative deviations of Gini and KS metrics and resulting stability scores for the top five and bottom five variables. Source: Authors' own elaboration.

Variable	Gini _{aβs}	Gini _{rel}	KS _{aβs}	KS _{rel}	Univariate Score	Bivariate Score	Final Score
Top 1	3.26	21,712.67	3.36	373.60	0.83	0.77	0.80
Top 2	1.60	-	2.04	-	0.72	0.86	0.79
Top 3	0.41	-	1.32	-	0.72	0.84	0.78
Top 4	3.10	-	2.70	-	0.68	0.86	0.77
Top 5	2.12	210.79	2.49	245.39	0.80	0.70	0.75
Bottom 1	0	-	0	-	0	0	0
Bottom 2	0.47	16.08	0.47	16.08	0	0.07	0.04
Bottom 3	2.76	16.25	1.79	14.07	0.03	0.05	0.04
Bottom 4	6.12	18.49	4.45	17.13	0.0039	0.08	0.04
Bottom 5	0.25	11.28	0.56	14.91	0	0.09	0.05

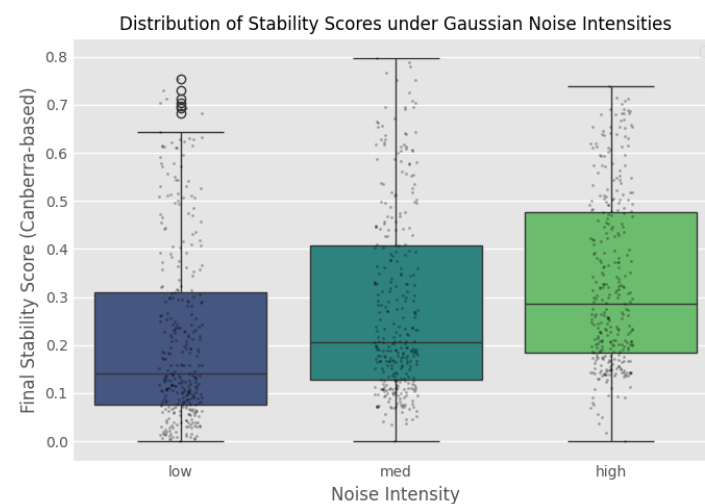


Figure 3. Distribution of Canberra-based final stability scores S_v under low, medium and high synthetic perturbation intensity ($p_{var} \in \{0.1, 0.3, 0.5\}$) applied to the reference dataset. Dots represent individual observations, while circles indicate statistical outliers identified by the boxplot criterion ($1.5 \times IQR$). Source: Authors' own elaboration.

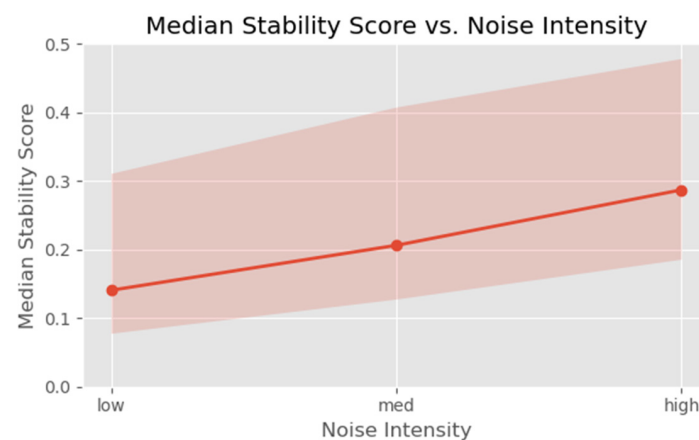


Figure 4. Median final stability score S_v as a function of synthetic perturbation intensity (p_{var}), demonstrating monotonic score escalation under increasing drift magnitude. The red line represents the median stability score across variables for each intensity level, and the shaded band indicates the interquartile range. Source: Authors' own elaboration.

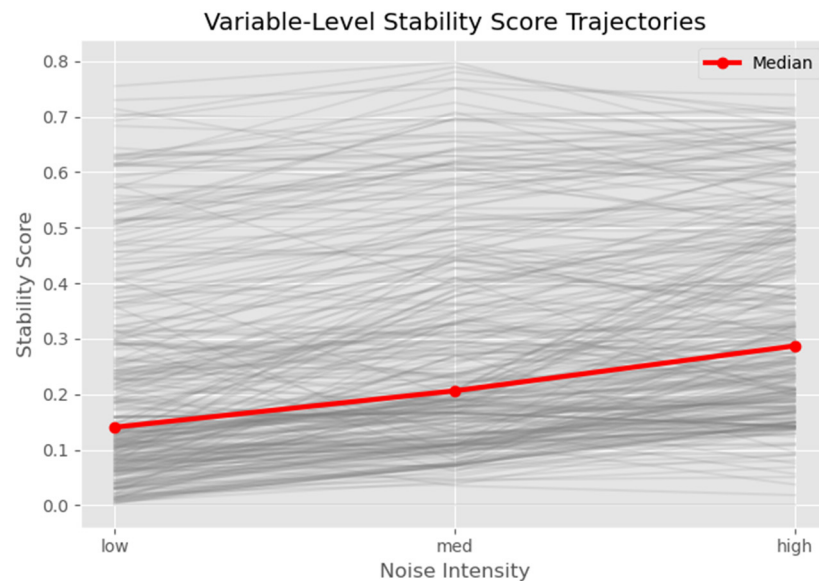


Figure 5. Variable-level trajectories of final stability score S_v across perturbation intensities, showing consistent drift amplification across features. Grey lines represent individual variable-level stability-level trajectories, and the red line denotes the median stability score at each intensity level. Source: Authors' own elaboration.

Collectively, these results confirm three desirable properties of the proposed approach:

- Sensitivity—stability scores increase systematically with perturbation magnitude.
- Proportionality—score escalation reflects ordered noise intensity rather than binary threshold effects.
- Consistency—variable-level trajectories demonstrate coherent drift amplification across features.

These findings provide empirical validation that the proposed monitoring framework produces interpretable and stable drift indicators under controlled synthetic distributional change.

3.6. Ablation Analysis of Monitoring Components

To assess the contribution of individual components within the proposed monitoring pipeline, an ablation-style analysis was conducted by examining drift scores derived from partial metric subsets (Table 11). Specifically, the effects of univariate-only and bivariate-only monitoring were compared against the fully aggregated stability score.

- Univariate-Dominant Drift—Variable Top 3 exhibits a high univariate score ($S_{uni} \approx 0.82$) but a comparatively moderate bivariate score ($S_{biv} \approx 0.56$), indicating drift driven primarily by marginal statistical changes. In a univariate-only configuration, this variable would be correctly prioritized, whereas bivariate-only monitoring would underrepresent its drift severity.
- Bivariate-Rescued Drift—In contrast, variable Top 4 exhibits a relatively low univariate score ($S_{uni} \approx 0.30$) but a high bivariate score ($S_{biv} \approx 0.85$). A monitoring system relying solely on univariate metrics would therefore fail to prioritize this feature, despite clear evidence of structural drift captured by dependency-based metrics.

The aggregated stability score (S_v) consistently demonstrates higher robustness by combining both perspectives. Variables affected either by marginal or relational drift are ranked more prominently under the combined score than under either ablated configuration alone. This confirms that the hierarchical aggregation of univariate and bivariate metrics

improves drift detection coverage and reduces the risk of false negatives arising from reliance on a single statistical view.

3.7. Operational Case Studies: Real-World Drift Validation

To demonstrate the framework's practical utility in identifying specific failure modes, we present two anonymized case studies derived from historical ETL monitoring results. These examples illustrate how the multi-component scoring mechanism successfully detects distinct types of structural and distributional inconsistencies that may not be identifiable through single-metric monitoring approaches.

Case Study 1: Silent Relationship Break (Correlation Disruption)

- **Scenario:** Variable Top 4 (Table 11) exhibited a relatively low univariate stability score ($S_{uni} \approx 0.31$), suggesting that its marginal distribution (mean, variance, dispersion) remained largely consistent with the reference dataset.
- **Detection:** In contrast, the bivariate stability score was substantially high ($S_{biv} \approx 0.86$), indicating significant changes in dependency structure relative to other variables.
- **Diagnosis:** Further inspection revealed that although the variable's individual values were within expected ranges, its relationships with key covariates had been disrupted. This pattern is characteristic of value shuffling or upstream transformation errors (e.g., incorrect join condition), where data remains syntactically valid but becomes semantically misaligned. A monitoring approach relying solely on univariate statistics (e.g., KS test or PSI) would likely classify the feature as stable. The hierarchical aggregation of univariate and bivariate components within the proposed framework correctly elevated the variable's overall stability score, prioritizing it for investigation.

Case Study 2: Structural Population Shift (Schema Evolution)

- **Scenario:** Variable Top 1 (Tables 3 and 4) underwent a substantial structural transformation between the reference and current ETL cycles. In the reference dataset, it appeared as a low-cardinality feature (4 unique values) with a near-zero central tendency. In the current ETL delivery, it transitioned into a high-cardinality continuous variable (1663 unique values) with a significantly shifted mean (1.55).
- **Detection:** The framework assigned near-maximum drift scores across both univariate ($S_{uni} \approx 0.86$) and bivariate ($S_{biv} \approx 0.86$) components, resulting in a high final stability score.
- **Diagnosis:** This behavior is consistent with a schema evolution event in which a previously categorical or flag-like variable is repurposed upstream to store continuous transactional information. The normalization strategy described in Section 2.6 ensured numerical stability despite the large relative increase in cardinality, while the aggregated stability score clearly indicated the need for schema review and downstream model assessment.

3.8. Summary of Findings

Across the reported analyses, the automated monitoring framework identifies and quantifies data drift between successive ETL cycles in a consistent and interpretable manner. Univariate metrics capture changes in distributional spread, bivariate metrics reveal changes in dependency structure, and Canberra-based aggregation produces bounded stability scores that support variable ranking and diagnostic investigation. Controlled noisification experiments confirm sensitivity and proportional response under synthetic perturbations, while operational case studies demonstrate that the approach surfaces both marginal and relational drift scenarios that may bypass conventional rule-based validation.

4. Discussion

The results demonstrate that the proposed statistical monitoring framework effectively characterizes both univariate and bivariate stability between a validated reference dataset and successive ETL outputs. Across multiple analytical stages, the framework consistently distinguishes stable variables from those exhibiting meaningful distributional or structural change, supporting its suitability as an automated, EDA-based monitoring component within recurring ETL pipelines.

At the baseline level, numerical variables in the reference dataset generally exhibited low dispersion, narrow interquartile ranges, and consistent central tendency. This provides a stable reference against which changes in subsequent ETL runs could be assessed. Variables with minimal baseline variability are particularly sensitive to distributional expansion, enabling the detection of early drift signals. In contrast, low-cardinality categorical variables tend to remain stable across ETL cycles, reflecting consistent encoding and schema preservation within the upstream process.

Bivariate analysis further indicates that dependency structures are largely preserved across ETL cycles. Correlation-based metrics and distributional indicators such as the Gini coefficient, KS statistic, and diversity index remain stable for most features, with only modest shifts observed among variables ranked as highly drifting. This distinction is important diagnostically: marginal distributions may evolve while inter-variable relationships remain intact, implying natural population drifts rather than structural corruption. Conversely, preserved marginals accompanied by relational breakdown may signal upstream join errors or semantic transformation issues.

Analysis of absolute and relative differences confirms that drift magnitude is unevenly distributed across features. Some features exhibit large absolute changes but moderate relative differences, while others show inflated relative deviations due to near-zero baseline values. This behavior is typical for ratio-based metrics and reinforces the need to combine absolute deviations, relative changes, and normalized, bounded aggregation rather than relying on a single indicator. Across all stages, categorical variables consistently show negligible changes in both marginal and relational metrics, supporting robust handling of discrete features within a unified monitoring workflow.

The framework is intentionally positioned as a model-agnostic and statistically grounded monitoring layer within the ETL pipeline. This aligns with broader data lifecycle management challenges in production machine learning systems, where upstream data validation is critical to downstream model reliability and governance [8]. Recent MLOps-oriented studies further emphasize monitoring and mitigation within production environments [20]. Design choices such as reliance on aggregated summary statistics, equal variable weighting, and normalized distance-based aggregation reflect a deliberate trade-off between interpretability, scalability, and general applicability. Rather than optimizing for a specific downstream model or business objective, the framework prioritizes early detection of distributional and structural inconsistencies that may indicate upstream data-quality issues, transformation errors, or evolving data-generating processes.

Taken together, these findings indicate that the framework functions effectively as an early-warning mechanism for detecting data drift, transformation inconsistencies, and emerging dependency changes while supporting root-cause analysis through metric-level audit outputs.

Importantly, the framework maintains interpretability at the individual metric level, allowing practitioners to trace elevated stability scores back to specific statistical deviations rather than treating drift as an opaque model output.

4.1. Comparison with Existing Approaches

Traditional ETL-monitoring and data-quality solutions rely primarily on rule-based validation, schema checks, and fixed thresholds [1–3]. While effective at detecting gross structural issues—such as missing columns or type mismatches—these approaches provide limited visibility into distributional or dependency-level changes.

In contrast, the machine learning literature offers advanced drift detection methods based on distributional distances such as Kullback–Leibler divergence, Jensen-Shannon divergence, and Wasserstein distance [14–17]. Although powerful, these techniques often require density estimation, assume continuous features, or impose computational costs that limit their practicality for large-scale, mixed-type ETL monitoring.

The proposed framework bridges this gap by integrating univariate and bivariate statistical profiling with a normalized Canberra-based scoring mechanism. Unlike many model-centric or distribution-level approaches, it operates exclusively on EDA-derived summary statistics, enabling lightweight deployment in production ETL workflows without requiring access to raw data streams or learned representations. This abstraction level supports heterogeneous feature types while capturing both marginal and relational changes, providing a more comprehensive view of ETL stability than single-metric or schema-only detectors.

Direct quantitative benchmarking against alternative drift detection is non-trivial in the absence of labeled drift events or a binary detection objective. In ETL monitoring contexts, the objective is not binary classification accuracy but interpretable prioritization and diagnostic transparency, which alters the evaluation paradigm relative to supervised drift detection frameworks. The proposed framework produces continuous, diagnostic stability scores rather than discrete drift decisions, and is designed for interpretability and operational monitoring rather than supervised detection accuracy. Consequently, baseline methods are discussed conceptually to highlight methodological differences, detection capabilities, and practical trade-offs relevant to ETL monitoring.

To contextualize the effectiveness of the proposed Canberra-based scoring, we contrast it with three established baseline methodologies commonly used in financial data monitoring: the Population Stability Index (PSI), the Kolmogorov–Smirnov (KS) test, and Wasserstein distance.

PSI is a widely used industry standard, particularly in credit risk monitoring. However, its dependence on binning can make it brittle in dynamic ETL environments. For sparse numerical features or high-cardinality categorical variables, PSI may yield unstable values due to empty or shifting bins. In contrast, the proposed framework handles cardinality shifts through explicit metrics (e.g., unique-value counts and diversity measures) and avoids reliance on fixed discretization.

Although the KS statistic is included as one component of the framework, its role differs from that of a standalone drift detector. In the proposed approach, the KS statistic contributes as a single component within a multi-metric aggregation and is never interpreted in isolation. As a univariate statistic, KS does not capture relationship drift—such as semantic breaks caused by joins or transformations—where marginal distributions remain similar but inter-variable dependencies change. The proposed framework addresses this limitation through explicit bivariate monitoring.

Finally, while the Wasserstein distance can be sensitive to geometric distributional shifts, its computational requirements and unbounded scale make it less suitable for routine ETL monitoring across large numbers of heterogeneous variables. The proposed Canberra-based aggregation normalizes inputs to a bounded $[0, 1]$ scale, improving interpretability and facilitating consistent prioritization across metrics with different units (e.g., currency amounts versus percentages), while keeping scoring computationally lightweight.

4.2. Strengths of Approach

A key strength of the framework lies in its layered design, combining baseline profiling, metric-level deviation diagnostics, and normalized aggregation. This enables both granular investigation and high-level stability ranking. Canberra-based normalization ensures proportional contribution across heterogeneous metric scales, preventing dominance by high-variance features and supporting consistent interpretation.

Operationally, the approach is scalable, reproducible, and integrates naturally into Python-based ETL ecosystems. Outputs such as variable-level scores and audit tables support efficient root-cause analysis of upstream transformation errors and data-quality issues.

Computational Complexity and Runtime Efficiency

The framework comprises two operational stages:

1. **Metric Extraction (ETL Layer)**—Extraction of univariate summary statistics scales as $O(NP)$, where N denotes observations and P denotes monitored variables. For rank-based metrics (e.g., Spearman correlation) or inequality measures requiring sorting, the dominant term is $O(P N \log N)$. These costs occur during standard EDA and are not unique to the monitoring framework.
2. **Drift Scoring (Monitoring Layer)**—Drift scoring operates only on precomputed summary statistics, decoupling monitoring latency from data volume. For each variable, scoring is $O(M)$, where M denotes the number of monitored metrics. Total report generation scales as $O(PM)$, independent of N .

Comparative Efficiency—in contrast, many distribution-level drift detection methods require reprocessing raw data and may range from $O(N \log N)$ to $O(N^2)$ depending on the implementation. By shifting computation to summary-statistic extraction (already required in ETL validation), the Canberra-based scoring introduces negligible overhead during monitoring and is well suited for large-scale, recurring ETL workflows.

4.3. Handling False Positives and False Negatives

Unlike supervised classification systems, where false positives (FP) and false negatives (FN) can be quantified against ground-truth labels, unsupervised data monitoring requires an operational and interpretative approach to error management.

Within the proposed framework, a false positive (Type I error) corresponds to a variable exhibiting a high stability score due to benign factors—such as seasonal effects or expected business growth—rather than unintended data quality degradation. The framework mitigates this risk through its ranking-based design. By prioritizing only the top- k drifting variables for review, the number of potential false alerts is bounded, reducing alert fatigue in high-dimensional ETL outputs. When an investigation confirms that observed drift reflects an accepted business change, this information can be incorporated during periodic reference dataset updates, preventing repeated flagging of expected evolution in subsequent monitoring cycles.

False Negatives (Type II error), where significant data issues remain undetected, pose a higher risk to downstream analytics. The framework reduces this risk through metric diversity and hierarchical aggregation. Changes that leave certain univariate statistics unchanged—for example, value shuffling that preserves mean and variance—are likely to affect relational or distributional properties captured by bivariate metrics such as correlations, Gini coefficients, or diversity indices. By jointly aggregating deviations across univariate and bivariate dimensions, the framework increases the likelihood that structurally meaningful drift is detected even when individual statistics remain stable, thereby reducing the probability of undetected drift events.

4.4. Extensibility of the Proposed Framework

The framework was designed with modularity and configurability in mind, allowing adaptation to a wide range of operational contexts beyond the experimental setup evaluated in this study. Its reliance on summary statistics, type-aware metric selection, and normalized aggregation enables straightforward extension along several dimensions, including data dimensionality, variable type heterogeneity, and monitoring frequency.

4.4.1. Extension to Higher-Dimensional Multivariate Data

The framework scales naturally to higher-dimensional datasets, as univariate and bivariate metrics are computed independently for each variable and selected variable pair. Drift quantification relies on deviations between summary statistics rather than joint density estimation or explicit multivariate modeling, thereby avoiding the curse of dimensionality that affects many distribution-based drift detection methods. Computational complexity grows linearly with the number of monitored variables and configured relationships, making the approach suitable for wide ETL tables commonly encountered in large-scale analytics pipelines.

4.4.2. Handling of Categorical and Mixed-Type Variables

The framework explicitly supports mixed data types, reflecting the heterogeneous nature of real-world ETL outputs. Numerical variables are monitored using descriptive statistics and correlation-based measures, while categorical and ordinal variables are assessed through frequency-based descriptors, diversity indices, Gini coefficients, Kolmogorov–Smirnov statistics, and hypothesis-based measures. Metrics that are undefined for a given variable type are retained for diagnostic transparency but are excluded from aggregations requiring finite values via missing-value propagation, enabling a unified workflow without separate pipelines or specialized encodings.

4.4.3. Adaptation to Alternative Monitoring Frequencies

Although the experimental evaluation presented in this study focuses on monthly ETL cycles, the proposed framework is not tied to a specific temporal resolution. The reference-current comparison paradigm applies equally to daily, weekly, quarterly, or event-driven data deliveries, provided that an appropriate benchmark dataset is defined. Monitoring frequency can be adjusted through updates to the reference dataset or rolling reference windows without modification of the core scoring logic.

4.5. Limitations

Several limitations should be acknowledged. First, Canberra-based ratios can become unstable when baseline values approach zero. The proposed framework mitigates this risk through explicit normalization, ϵ -stabilization, and bounded aggregation, as described in Section 2.6. Nevertheless, interpretation of large relative deviations remains context dependent and benefits from complementary diagnostic metrics.

Second, the proposed framework adopts a fixed reference window strategy. While this design enhances sensitivity to cumulative and gradual distributional drift, it assumes that the reference dataset remains representative over the monitoring horizon. In operational settings, this requires periodic review and controlled updates of the benchmark following major schema revisions. Additionally, this fixed-window approach does not explicitly model strong seasonal effects (e.g., end-of-year transaction spikes), which may manifest as distributional drift relative to a static benchmark. Differentiating expected seasonal variation from anomalous data quality issues currently requires human interpretation.

Alternative windowing strategies, such as rolling or adaptive reference windows, represent a natural direction for future extensions.

Third, the current aggregation strategy assigns equal weight to all variables by design, prioritizing transparency and model-agnostic applicability. While this may underrepresent the operational importance of certain features in specialized contexts, it ensures consistent drift detection across heterogeneous datasets and avoids reliance on proprietary business knowledge or downstream models.

Fourth, the framework aggregates heterogeneous metric deviations into a single normalized score. Although this provides an interpretable and computationally efficient summary, collapsing multiple indicators into a single scalar value may not capture all aspects of distributional change simultaneously. This motivates the ensemble-based and multidimensional extensions discussed in future work.

Finally, while hypothesis-test outputs are computed and reported diagnostically, they are not currently incorporated into automated alerting or decision logic. Distinguishing systematic drift from random fluctuations in fully autonomous monitoring scenarios therefore remains an area for further development.

4.6. Future Work

Future research may focus on adaptive weighting, multidimensional drift characterization, inferential validation, and automated alerting mechanisms. These extensions would allow the framework to evolve from a descriptive monitoring tool into a more decision-oriented component for large-scale production environments.

4.6.1. Adaptive Weighting Schemes

A natural extension of the proposed framework involves the incorporation of adaptive weighting schemes in the aggregation of variable-level drift scores. Rather than assigning equal importance to all monitored variables, weights could be derived from historical volatility profiles, allowing variables with inherently stable behavior to be penalized more strongly when drift occurs. In addition, domain-specific criteria such as business criticality, regulatory relevance, or downstream usage could be incorporated to prioritize variables whose instability poses higher operational risk.

Further extensions could integrate information from downstream predictive models, such as feature importance measures or sensitivity analyses, enabling drift scores to reflect not only statistical deviation but also potential impact on model performance. Within the current framework, such weighting schemes could be introduced as a configurable layer applied during score aggregation, without altering the underlying metric computation. This would allow stability assessments to be more closely aligned with real-world risk and decision-making priorities. Such extensions align with emerging MLOps taxonomies that emphasize modular, decoupled validation layers within production AI systems [21].

4.6.2. Integration of Alternative Drift Metrics

The current implementation relies on a normalized Canberra-based distance to aggregate deviations across heterogeneous metrics. While this provides an interpretable and scale-robust indicator, future work could extend the framework by incorporating complementary drift measures such as Wasserstein distance or Jensen–Shannon divergence. These metrics capture different aspects of distributional change, including shifts in scale, shape, support, and categorical structure, which may not be fully reflected by a single distance measure.

An ensemble of normalized drift metrics could be constructed by computing each distance independently and combining them through averaging, weighted aggregation, or diagnostic comparison. Rather than replacing the Canberra-based score, such an ensemble

would provide a multidimensional characterization of drift, enabling practitioners to distinguish between different drift patterns and supporting more nuanced root-cause analysis.

To further address the limitations of linear correlation measures, future iterations could incorporate non-linear dependency metrics for a targeted subset of high-priority variables. Information-theoretic measures, such as Mutual Information (MI) or Distance Correlation, could be computed when linear drift scores exceed a preliminary warning threshold. This hybrid approach would balance the need for detecting deep structural drift and non-linear interactions with the low-latency requirements of production ETL pipelines.

4.6.3. Hypothesis-Based Drift Validation

While the current framework relies on descriptive statistics and distance-based indicators to quantify distributional change, future work could enhance its inferential foundation through the selective integration of formal hypothesis testing. Such tests would provide statistical evidence regarding whether observed deviations are consistent with random variation or indicative of systematic drift, complementing rather than replacing the existing metric-based diagnostics.

For continuous variables, goodness-of-fit and distributional equality tests such as the Cramér-von Mises or Anderson–Darling tests could be employed to assess changes in overall distributional shape, with particular sensitivity to differences in the tails. These tests are well suited for detecting subtle but persistent deviations that may not strongly affect central tendency measures yet carry operational significance in financial data contexts. In parallel, location-focused tests such as the two-sample *t*-test or the nonparametric Mann–Whitney U test could be used to identify shifts in central tendency, depending on distributional assumptions and sample characteristics.

Changes in dispersion could be evaluated using variance-focused tests such as Levene’s or Bartlett’s tests, allowing the framework to distinguish between shifts driven by increased volatility and those arising from location or shape changes. This separation is particularly relevant for monitoring ETL pipelines, where upstream process changes may affect variability without substantially altering means or medians.

Within the proposed framework, hypothesis tests would be applied at the variable level and their outcomes incorporated as auxiliary signals alongside descriptive drift scores. Rather than enforcing binary accept-reject decisions, test statistics or *p*-values could be reported as contextual indicators, supporting informed interpretation and prioritization. This hybrid approach would preserve operational sensitivity while strengthening statistical rigor, helping practitioners balance early drift detection against false positive control in continuous monitoring environments.

4.6.4. Adaptive Thresholding and Alerting

At present, the framework reports continuous stability scores without automated alerting. Future work could extend this functionality by introducing adaptive thresholding mechanisms derived from historical behavior. Thresholds could be defined using rolling baselines, percentile-based rules, or statistical process control techniques such as control charts, allowing the system to adapt to non-stationary environments.

By integrating such mechanisms, the framework could evolve into a semi-autonomous drift detection and prioritization system, capable of flagging anomalous behavior while accounting for expected variability. This would be particularly valuable in large-scale production environments, where manual monitoring is infeasible and monitoring resources must be allocated efficiently.

4.6.5. Handling Seasonality and Non-Stationarity

Finally, to address the limitations of the fixed reference window regarding seasonal effects, future research will explore “Seasonal Reference Windows” (e.g., comparing the current month’s output against the same month from the previous year). Additionally, for high-frequency data, integrating time-series decomposition techniques—such as Seasonal Trend decomposition using Loess (STL)—would allow the framework to isolate true distributional drift from expected cyclic variations. This extension is crucial for distinguishing between regular business cycles (e.g., end-of-year transaction spikes) and genuine data quality degradation.

5. Conclusions

This study demonstrated that distributional and structural drift in recurring ETL outputs can be reliably detected using a summary-statistic-based monitoring framework combined with a normalized Canberra aggregation scheme. Across real-world ETL comparisons and controlled perturbation experiments, the framework consistently separated stable variables from those exhibiting meaningful change. Variables affected by variance inflation, shifts in cardinality, or altered dependency structures achieved substantially higher stability scores than unaffected features, confirming discriminative ranking capability.

Synthetic noisification experiments further validated sensitivity under controlled conditions. Increasing perturbation intensity produced monotonic increases in aggregated drift scores and systematic widening of score dispersion. This confirms that the proposed scoring mechanism responds proportionally to progressive degradation of statistical structure rather than triggering arbitrary or threshold-driven alerts.

The primary methodological contributions of this work are fourfold:

1. A hierarchical aggregation strategy that integrates univariate and bivariate monitoring metrics into a unified, bounded stability score while preventing dominance by any single statistical dimension.
2. A normalized Canberra-based formulation tailored to heterogeneous ETL summary statistics, enabling scale-robust comparison across metrics with different units, magnitudes, and baseline variability.
3. A fixed-reference window monitoring design that mitigates gradual drift masking and preserves sensitivity to cumulative structural change in batch-oriented ETL workflows.
4. An operationally decoupled architecture in which computationally intensive metric extraction is separated from lightweight drift scoring, allowing monitoring latency to remain independent of dataset size and suitable for large-scale production environments.

Together, these contributions provide a reproducible and interpretable alternative to raw-distribution or model-centric drift detectors, particularly in environments where access to full historical data streams or labeled drift events is impractical.

Several limitations remain. Relative differences may become unstable when baseline values approach zero, requiring careful stabilization and transparent reporting. The current aggregation assigns equal weight to all variables and does not incorporate automated alert thresholds or formal hypothesis testing. Future work will explore adaptive weighting strategies, ensemble drift metrics, statistically grounded alarm calibration, and longitudinal extensions incorporating rolling or adaptive reference strategies.

Beyond the financial ETL context examined here, the framework is domain-agnostic and applicable wherever recurring data deliveries must be validated against a stable reference. Potential applications include industrial measurement systems, scientific simulation pipelines, high-dimensional biomedical data integration, and large-scale socioeconomic databases. In each case, the ability to detect structural and distributional drift using lightweight, interpretable summaries offers practical advantages over raw-data-dependent approaches.

In summary, this work establishes a scalable and transparent methodology for automated ETL stability monitoring. By combining descriptive statistical profiling with normalized hierarchical drift aggregation, it provides actionable diagnostics while preserving computational efficiency and extensibility across diverse operational domains.

Author Contributions: Conceptualization, K.P. and A.E.; methodology, K.P. and A.E.; software, I.G., A.K. and K.P.; validation, K.P., I.G. and A.K.; formal analysis, K.P., I.G. and A.K.; investigation, K.P.; resources, I.G. and A.K.; data curation, I.G. and A.K.; writing—original draft preparation, K.P.; writing—review and editing, K.P., I.G. and A.E.; visualization, K.P.; supervision, A.E.; project administration, K.P.; funding acquisition, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This study is financed by the European Union—NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0005.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it did not involve any sensitive or personal data and/or invasive procedures. This research was conducted in accordance with the regulations of the Faculty of Automatics, Technical University of Sofia. The specific approval details are maintained by the department.

Informed Consent Statement: Informed consent for participation was obtained from all subjects involved in the study.

Data Availability Statement: Data used in this study is unavailable due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Firmani, D.; Mecella, M.; Scannapieco, M.; Batini, C. On the Meaningfulness of “Big Data Quality”. *Data Sci. Eng.* **2016**, *1*, 6–20. [\[CrossRef\]](#)
2. Abedjan, Z.; Chu, X.; Deng, D.; Fernández, R.C. Detecting data errors: Where are we and what needs to be done? *Proc. VLDB Endow.* **2016**, *9*, 993–1004. [\[CrossRef\]](#)
3. Schelter, S.; Lange, D.; Schmidt, P.; Celikel, M.; Biessmann, F.; Grafberger, A. Automating Large-Scale Data Quality Verification. *Proc. VLDB Endow.* **2018**, *11*, 1781–1794. [\[CrossRef\]](#)
4. Janssen, M.; Brous, P.; Estevez, E.; Barbosa, L.S.; Janowski, T. Data Governance: Organizing Data for Trustworthy Artificial Intelligence. *Gov. Inf. Q.* **2020**, *37*, 101493. [\[CrossRef\]](#)
5. Grafberger, S.; Groth, P.; Stoyanovich, J.; Schelter, S. Data distribution debugging in machine learning pipelines. *VLDB J.* **2022**, *31*, 1103–1126. [\[CrossRef\]](#)
6. Breck, E.; Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data Validation for Machine Learning. In Proceedings of the MLSys, Stanford, CA, USA, 31 March–2 April 2019.
7. Caveness, E.; Cukierski, W.; Park, J.; Polyzotis, N.; Zinkevich, M. Data Analysis and Validation in Continuous ML Pipelines (TensorFlow Data Validation). In Proceedings of the ACM SIGMOD, Virtual, 14–19 June 2020. [\[CrossRef\]](#)
8. Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data Lifecycle Challenges in Production Machine Learning: A Survey. *SIGMOD Rec.* **2018**, *47*, 17–28. [\[CrossRef\]](#)
9. Webb, G.I.; Hyde, R.; Cao, H.; Nguyen, H.L.; Petitjean, F. Characterizing concept drift. *Data Min. Knowl. Discov.* **2017**, *30*, 964–994. [\[CrossRef\]](#)
10. Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; Zhang, G. Learning under concept drift: A review. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 2346–2363. [\[CrossRef\]](#)
11. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 44. [\[CrossRef\]](#)
12. Vaquet, V.; Hinder, F.; Hammer, B. One or two things we know about concept drift—A survey on monitoring in evolving environments. *Front. Artif. Intell.* **2024**, *7*, 1330257. [\[CrossRef\]](#)
13. Rabanser, S.; Günnemann, S.; Lipton, Z.C. Failing loudly: An empirical study of methods for detecting dataset shift. In Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019. [\[CrossRef\]](#)
14. Kullback, S.; Leibler, R.A. On information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [\[CrossRef\]](#)
15. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [\[CrossRef\]](#)
16. Villani, C. The Wasserstein distances. In *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009. [\[CrossRef\]](#)

17. Peyré, G.; Cuturi, M. Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [[CrossRef](#)]
18. Massey, F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. [[CrossRef](#)]
19. Aggarwal, C.C.; Hinneburg, A.; Keim, D.A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory—ICDT 2001*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 420–434. [[CrossRef](#)]
20. Hussain, N. Data drift detection and mitigation: A comprehensive MLOps approach for real time systems. *Int. J. Sci. Res. Arch.* **2024**, *12*, 3127–3139. [[CrossRef](#)]
21. Testi, M.; Ballabio, M.; Frontoni, E.; Iannello, G.; Moccia, S.; Soda, P.; Vessio, G. MLOps: A Taxonomy and a Methodology. *IEEE Access* **2022**, *10*, 63606–63618. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.