Assessing and Visualizing Completeness, Co-Coverage, and Scalability in Multivariate Time-Series Data

Long Vu*, Madeline Frank*, Honghui Xu*, Sisi Chen†, Tu N. Nguyen*,
Selena He*, Bobin Deng*, Kun Suo*
*College of Computing and Software Engineering, Kennesaw State University, Marietta, Georgia, USA
†College of Health Professions, Mercer University, Macon, Georgia, USA
{lvu6, mfrank64}@students.kennesaw.edu, hxu10@kennesaw.edu, chen_s@mercer.edu,
{tu.nguyen, she4, bdeng2, ksuo}@kennesaw.edu

***Abstract*—Assessing data quality in multivariate time-series datasets is crucial for reliable analysis, particularly when dealing with missing values, inconsistent feature availability, and massive records in large-scale edge computing and IoT clusters. Existing methods often fall short of capturing intricate patterns of missingness and co-coverage, restricting the capacity to make well-informed decisions regarding the usability of the data. In order to systematically extract reliable data segments, this paper presents a comprehensive framework that combines a heuristic model with temporal coverage, period-specific missingness, and co-coverage metrics. By integrating these metrics with visualizations such as temporal coverage heatmaps and parallel coordinates plots, the framework reveals complex patterns of missingness while supporting human involvement in validating data subsets. Our approach effectively balances automation with expert judgment, enhancing the interpretability of data quality assessments. The findings show that the proposed methods satisfy the design specifications for revealing patterns, quantifying missingness impact, measuring feature availability, guiding feature selection, and facilitating scalable, multi-scale data summarization. The framework offers a solid way to improve the quality of data in multivariate time-series analysis, opening the door to more precise and trustworthy insights for assessing data gathered from edge computing infrastructures and large-scale, heterogeneous IoT deployments, where data consistency and completeness are frequently very variable.**

***Index Terms*—Missingness, Co-Coverage, Data Quality, IoT**

I. Introduction

Multivariate time series data are common in large-scale edge computing and IoT clusters. For example, in modern smart grids, a large number of edge devices (such as smart meters, substation sensors, and grid control terminals) continuously collect multiple data points such as voltage (V), current (A), power factor (PF), etc. Missing data is a key challenge which directly affects the reliability of downstream analysis. Thus, accurately assessing the extent and pattern of missing data is essential. Visualizing missing data is an effective means to address this challenge. Effective visualization can not only reveal the pattern of missing data, but also assess its impact on the relationship between time series features, thereby guiding the selection of data interpolation or preprocessing strategies [1]. In multivariate time series analysis, these visualization methods must deal with standard coverage in the time dimension. Customized visualization techniques can highlight data inconsistencies, helping domain experts to extract reliable insights in incomplete data environments [2].

However in large-scale datasets generated by IoT devices and automated systems, existing methods often struggle to provide comprehensive ways to evaluate and visualize data missingness [2]. This challenge is particularly prominent in fields that rely on multivariate time series data for complex system or environmental monitoring and modeling. Such data are often generated by heterogeneous sensor networks or automated systems and are therefore prone to missing or inconsistent data [3]. In addition, the complexity of data quality assessment increases further with the increasing size of time series datasets and the lack of unified standards across different data sources. There are currently no automated techniques that can successfully adjust to the complexity and particular requirements of multivariate time series data, despite the fact that there are some visualization tools available for analyzing missing data [4].

Compared to traditional efforts on visualizing missing data and data quality often falls short of capturing complex patterns of missingness and co-coverage in large-scale datasets, to the best of our knowledge, our work is the first to explore integrated methodology at both temporal and cross-feature levels in multivariate time-series settings. By addressing challenges commonly found in data from IoT deployments and edge computing infrastructures, our contributions fill an important gap and lay the foundation for future comparative and application-driven evaluations in assessing and extracting reliable segments from multivariate time-series data with heterogeneous completeness and feature co-coverage.

A. Existing Challenges

Our work analyzes necessity of multivariate time series datasets collected from various sources to gain insights into temporal patterns, correlations, and anomalies. These datasets are typically collected from different locations with contexts and variables relevant to their respective environments and ap-

TABLE I: Design Goals for Addressing Challenges in Multivariate Time Series Datasets

Challenge	Goal	Explanation
P1	(G1) Reveal Missingness Patterns and Trends	Identify recurring gaps, trends, and correlations with recorded values
	(G2) Quantify Missingness Impact	Assess the severity and distribution of missingness to determine its effect
P2	(G3) Measure Feature Availability Consistency	Ensure reliable co-coverage between features for meaningful analysis
	(G4) Guide Feature Selection	Highlight strong and weak co-coverage to aid in selecting robust feature sets
P3	(G5) Support Multi-Scale Data Summarization	Provide both high-level overviews and detailed drill-downs
	(G6) Enable Efficient Interaction	Ensure smooth navigation, filtering, and comparative analysis across datasets

plications. After our analysis, multivariate time series datasets face the following challenges:

P1: Large Amounts of Missing Data. When examining these datasets, we observed that missing data is a widespread problem. Many factors, including sensor failures, variants in deployment, maintenance practices, and differences in data collection protocols, could contribute to this. For instance, researcher have reveal that many datasets that cover long period of time often reveal significant data gaps.

P2: Multivariate Co-Coverage. Normally, there exist complex relationships between multiple variables and their temporal dynamics in multivariate time series datasets. For example, in industrial IoT devices, data such as bearing temperature (°C), vibration frequency (Hz), noise level (dB), current (A), and pressure (Pa) are highly coupled. However, due to the heterogeneity of these datasets, they differ in measured variables, temporal resolution, and exhibit different degrees of missing coverage. All these differences might affect the ability to draw reliable conclusions or hypotheses in the datasets.

P3: Scalability of Detection and Verification. Multivariate time series datasets are often very large. Their large size also increases the complexity of analyzing their data integrity. For example, datasets across multiple sources or large sensor networks may contain thousands of data points, making manual verification and control of data quality impractical. Often, an effective solution for detecting missing data may not necessarily work for other datasets.

B. Design Requirements

In order to effectively analyze multivariate time series data with missing values, inconsistent feature availability, and large-scale records, powerful visualization methods are essential. To achieve this goal, we define key visualization goals to guide our assessment of missingness, common coverage, and scalability of the data, as shown in Table I. To summarize, the contributions in our work are listed as follows:

- **Comprehensive Metrics and Heuristic Model:** We designed a heuristic model combined with a set of metrics to systematically extract reliable data segments for consistent and high-quality subset analysis.
- **Effective Visualizations for Missingness and Temporal Co-Coverage:** Our visualizations make it possible for users to interactively examine the quality of the data and make well-informed choices regarding feature selection and usability by exposing complex patterns of missingness and co-coverage.
- **Scalable Framework for Multivariate Time-Series Analysis:** The proposed framework solves the problems caused by big, high-dimensional datasets by supporting multi-scale data summarization and improving scalability.
- **Human-Centric Decision Support:** Through integrating visualization, we achieve an effective balance between human involvement and automation, which helps experts interpret patterns and validate data more reliably.

This paper is organized as follows: In Section I, we discussed motivations and design requirements for assessing data quality in multivariate time-series datasets. Section II reviews existing approaches in visualizing missingness and scalability challenges. Section III introduces metrics and heuristic model, along with visualizations (i.e., temporal coverage heatmaps, parallel coordinates plots, etc.). Section IV demonstrates the effectiveness of our approach in extracting reliable data and addressing the design requirements, discussing the strengths and limitations of the methods. Section V summarizes the contributions and outlines future enhancements for interactivity, scalability, and automated imputation strategies.

II. Related Work

A. Data Quality in Information Visualization

Across the majority of analytics considering data and analytics as a unified construct for information visualization in different domains, data is the foundational element of the greatest quality concern. Xie et al. describe the issue of including variable data quality in multivariate visualizations while sustaining synergy integrity. Tierney et al. identify a seamless solution to the problem of missing data by automating the bordering data cleaning workflows. In [5], Kandel et al. report the combination of context-aware visualization and automatic anomaly detection to facilitate the scalable visualization of errors in large datasets. Arbesser et al. [6] employ metadata and hierarchical aggregation to show the plausibility of a time series, thereby confirming the plausibility of thousands of series. Gschwandtner et al. [3] has built upon this work to enable interactive computer-human time series analysis by correlating critical automated inspection with interactive visualization. These works together show how the problem of data quality, whether multivariate missing intervals, time series, or automation at scale, is enhanced by visualization as a direct integration of computational and human inputs.

B. Visualizing Missing Data

A variety of approaches have been developed to visually represent missing values. One brilliant showcase of Visualization and Imputation of Missing Values (VIM) R package is Templ et al. [7], where the authors employed visual elements such as histograms, scatterplots, and parallel coordinates to display

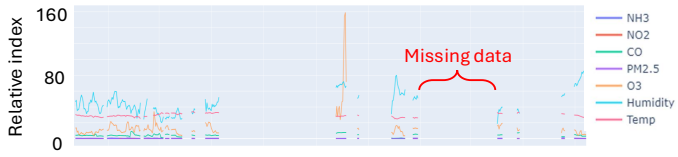

Fig. 1: Basic Data Visualizations for Time-Series Datasets

missingness attributes and assist in analysis. Another example is the work from Fernstad et al. [1], which outlines the MissiG system, a glyph-based application that integrates elements of multi-dimensional bar and histogram charting. Some other early tools, Missing Are Now Equally Treated (MANET) [8], incorporated missingness into charts, which makes data gaps more visible. More recent innovations, such as KYE [3], provide interactive exploration of missing values in time series and offer dynamic visualizations such as time series heat maps. The [9] naniar package brings a tidy approach to the integration of workflows for analyzing missing data. Simplistic techniques such as lasagna plots [10], missingness maps [11], and other tools, highlight the importance of flexible, simple approaches to assist in the analysis of incomplete data [4].

C. Missingness as a Feature

Beyond regular visualization, missing data has become a vital aspect in data analysis. By systematically validating missing data, meaningful structure problems such as systematic biases in survey responses or technical issues in data collection could be revealed. For instance, key patterns identified by Fernstad et al. [12], such as Amount Missing (AM), Joint Missingness (JM), and Conditional Missingness (CM), are designed to estimate unknowns of the collected data. MANET System [8], along with other preliminary methodologies, assessed missing parameters and the potential dependencies existing between the variable arrays. On the other hand, systems such as KYE [3] and MissVisG [13] offer more modern solutions, providing mixed interactive visualizations of missingness as well as employing glyph-based techniques to analyze missing data, whether structured or unstructured. These approaches to data, especially pertaining to missing responses in forms during health studies, offer more than just visualization. The very absence of response holds an underlying condition and trends that are valuable to health researchers. It is a misconception to classify missing pieces of data as a lack or deficit. With tools like MissVisG, we come to resolve a deficiency in understanding data, along with the potential systematic hidden in plain sight. However, tools like [3] still fall short in addressing the temporal granularity otherwise required with multivariate time series data.

III. Design and Implementation

Figure 1 shows the basic visualizations typically used to inspect time-series data. We can observe that if there exist a data missing inside, which is common for data collected from automated systems, but usually fall short in showing the severity of the missing data or the reliability of the remaining data. In order to evaluate missingness's effect on data quality, we first examine techniques for identifying, quantifying, and visualizing it. Then to assess how consistently features are recorded together and guarantee reliable multivariate relationships, we examine co-coverage analysis. Finally, we address scalability by introducing an automated framework that standardizes, validates, and visualizes large-scale datasets, aiming at fulfilling design requirements. The source code and raw data in this paper are available at https://github.com/longthangvu/ts-missingness-visuals.

A. Assessing Missingness for Time Series Data

In time-series datasets, missing data can skew statistical analyses, obscure trends, thus reduce model reliability. On the other hand, missingness in multivariate datasets can appear as discrete gaps or as systematic patterns that impact several variables at once. Due to these reasons, identifying and quantifying missingness is essential to ensure downstream analyses and visualizations are built on a reliable support.

1) Identifying Missingness Patterns and Trends: Missing data often follow certain patterns that affect data integrity, including: *(1) Variable-Specific Trends*: Some variables may consistently have a higher percentage of missing data than others; *(2) Simultaneous Missingness*: Multiple variables may experience gaps at the same time, which lowers their combined usability; *(3) Temporal Patterns*: Seasonal analysis may be impacted by missing data that corresponds with particular time periods (such as particular months or seasons); *(4) Correlations with Other Variables:* Missingness in one feature may be linked to recorded values in another, revealing systematic biases in data collection. Analyzing the distribution of missing values over time is essential to recognize systematic trends for better planning of preprocessing strategies.

Fig. 2: Temporal Coverage Heatmap for One Dataset

2) Quantifying Missingness: To assess the severity and distribution of missing data, we propose the following key metrics to provide a structured way to evaluate dataset completeness: (1) *Gap Length Statistics:* Measures the duration of consecutive missing values. For example, mean gap length provides insight into whether missingness is sporadic or in extended periods, which is problematic for time-series. (2) *Temporal Coverage:* Calculates how much of a given period contains valid data. Figure 2 visualizes temporal coverage by showing the percentage of valid data over time for multiple variables, with darker shades indicating higher coverage and lighter shades or gaps indicating missing data. This makes it easier to distinguish between times when data is consistently present and times when there are notable gaps across various features. (3) *Period-Specific Missingness:* Determines the proportion of missing data within specific time frames (e.g., months), which

can be calculated by aggregating coverage over defined time frames. This helps assess whether certain periods have enough data for meaningful trend analysis.

B. Understanding the Data Co-Coverage

In multivariate time-series datasets, the consistency of feature availability, *co-coverage*, is critical for meaningful analysis. It is difficult to make trustworthy correlations between features when they have missing values at different times. Co-coverage analysis guarantees that there are enough overlapping data points to examine feature interdependencies. However, current missingness visualization techniques are limited in their ability to display co-coverages between features, even though they are effective at illustrating temporal patterns. It doesn't specify whether various variables were recorded concurrently or separately; it just shows the total amount of data available. Consequently, it lacks insights into feature relationships and co-occurrence patterns, which are essential for multivariate analyses that depend on reliable co-coverage for accurate modeling and inference.

1) Measuring Feature Co-Coverage: To assess co-coverage, we define a *co-coverage matrix* that quantifies the percentage of time steps where each feature pair is simultaneously recorded. For two features F_i and F_j, co-coverage at time t is represented by:

$$r_{i,j,t} = \begin{cases} 1 & \text{if both } F_i \text{ and } F_j \text{ are recorded at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

We construct the overall co-coverage matrix A for a given time period p for which each element is calculated as: $A_{i,j} = \frac{\sum_{t=1}^{T} r_{i,j,t}}{T}$, where T is the total number of time steps in p. A high co-coverage value indicates strong joint availability of features, making them more suitable for multivariate analysis.

While this matrix can be visualized by simple heatmaps, we propose using parallel coordinates plot, which can offer a more intuitive understanding of co-coverage patterns by highlighting how shared data proportions change across features, as shown in Figure 3. The figure visualizes co-coverages between features by showing parallel lines that indicate the proportion of valid data points shared between pairs of variables. High co-coverage is indicated by lines that stay at similar levels across features, whereas lines that diverge show differences in the availability of data for different variables.

2) Identifying Reliable Feature Relationships: Finding trustworthy co-coverage relationships is essential because it helps researchers identify feature availability inconsistencies, evaluate the suitability of datasets for multivariate modeling, and direct feature selection. In order to prevent models from being trained on skewed or incomplete data, it is helpful to identify features that are commonly missing together. Additionally, by assessing co-coverage, researchers can determine whether a dataset offers sufficient joint observations for reliable multivariate modeling. Researchers can improve model performance and reliability by giving priority to features with high co-coverage, guaranteeing that analyses are carried out on statistically significant and well-represented data.

C. Heuristic Filtering Model

To extract high-quality segments of multivariate time-series data, we implement a lightweight heuristic filtering model based on two criteria: (i) per-feature temporal coverage, and (ii) joint co-coverage across feature combinations. Let $F = \{f_1, f_2, \ldots, f_n\}$ be the set of features and T the full timeline partitioned into fixed-size intervals. For each interval $t \in T$, we compute:

- **Feature coverage:** the proportion of non-missing values per feature f_i within t.
- **Pairwise co-coverage:** for each pair (f_i, f_j), the fraction of timestamps in t where both features are present.

An interval t is retained only if:

1) coverage$(f_i, t) \geq \theta_{\text{feat}}$ for all selected f_i
2) co_coverage$(f_i, f_j, t) \geq \theta_{\text{joint}}$ for all (f_i, f_j) pairs

Here, θ_{feat} and θ_{joint} are tunable thresholds that determine the minimum acceptable completeness. By default, we use $\theta_{\text{feat}} = 0.8$ and $\theta_{\text{joint}} = 0.7$, but these can be modified to balance strictness with data retention or to reflect domain-specific reliability requirements.

This produces quick, comprehensible results suitable for big datasets while avoiding the complexity of probabilistic filtering or learning-based imputation, generalizes effectively across domains with different sampling patterns and quality.

IV. Preliminary Result of Data Quality and Reliability Analysis

In the following section, we distinguish between the visual components of our proposed framework (Sections III) and the visualizations used to summarize and evaluate its effectiveness. While Figures 2, 3 and 4 are part of the visual toolkit designed to support the six design goals, Figures 5 and 6 serve to demonstrate the practical outcomes of applying our method across diverse datasets. Importantly, since each visualization contributes to different aspects of the data quality assessment, they are intended to be used together. The design goals outlined in Table I are fulfilled through the combined use of these views, rather than by any individual figure in isolation.

A. Datasets

We use open-source indoor air quality as an example to assess the data quality in multivariate time series datasets. Table II provides a summary of datasets, including the number of features, total data rows, average sampling frequency, and recording duration. These six datasets collect indoor air from multiple locations around the world, including Pune (India), Merida (Mexico), Brindisi (Italy), central Sweden, and two locations in California (a detached house and a low-income apartment). These datasets deployed different sensors and methods to monitor a variety of pollutants such as $PM_{2.5}$, NO_2, NH_3, CO, O_3, CO_2, as well as environmental variables like temperature and humidity. The difficulties of widespread, low-cost sensor deployment and irregular data collection methods are reflected in the data collection period's location-specific variations, temporal resolutions, and missing levels.

TABLE II: Summary of Datasets (*aggregated from multiple files within the same location)

Dataset Group	Total Rows	Columns (avg±std)	Duration (days)	Sampling Rate
India	173,465	9.0 ± 0.0	603.0 ± 0.0	every minute
Mexico	7,406	10.0 ± 0.0	367.0 ± 0.0	hourly
Sweden*	498,672	7.0 ± 0.0	1370.5 ± 0.5	every 10 minute
Calihome*	672,510	31.1 ± 3.9	6.1 ± 0.5	every minute
Caliapt*	228,083	50.1 ± 10.5	6.3 ± 0.8	every minute
Italy*	28,177,936	18.0 ± 0.0	348.3 ± 2.4	every 2 second

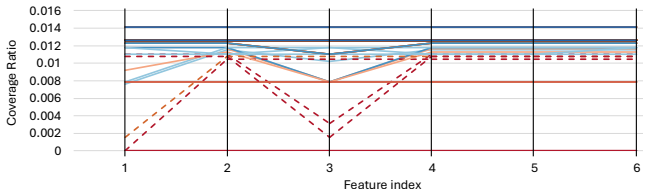

Fig. 3: Parallel Coordinates Plot of Average Co-Coverage Values for Features Across Datasets

B. Data Completeness and Coverage

To extract reliable portions of the data, we use a heuristic model in conjunction with the proposed metrics, which include temporal coverage, period-specific missingness, and co-coverage. Only high-quality, consistent subsets are retained for analysis thanks to this methodical filtering out features and time periods with inadequate data coverage. Using our suggested missingness quantification metrics, Figure 4 shows the temporal coverage heatmap for all six datasets, emphasizing the times when the data is complete and when it is missing. A dataset is represented by each row, and the shading's intensity indicates the proportion of reliable data that was gathered during particular time periods. The figure illustrates how some datasets, like "India," have large and irregular gaps, while others, like "Sweden" and "Mexico," maintain greater data coverage over longer periods of time. By highlighting missingness patterns and supporting multi-scale summarization, this visualization fulfills design goals G1 and G5, which allows users to quickly identify periods that are reliable and those that are not.

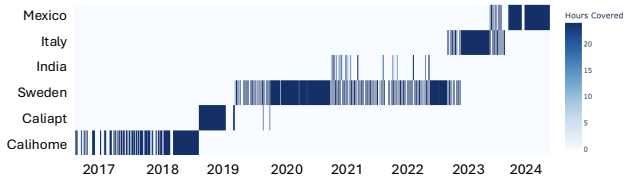

Fig. 4: Temporal Coverage of Multiple Datasets Over Time

C. Data Missingness and Usability Analysis

The parallel coordinates plot is useful for identifying feature pairs that have consistently low co-coverage across datasets, as shown in Figure 3. Since their combined sparsity lowers the effective sample size for multivariate modeling, such features can actually be deprioritized during model development or preprocessing in practice. This visualization directly supports G4 by guiding analysts toward feature subsets with strong mutual availability.

Baseline missingness and usable data percentages are compared between datasets in Figure 5. While the bars labeled "Percentage Usable Data (¿80%)" indicate data that has been filtered using our heuristic model based on missingness thresholds, the bars labeled "Percentage Recorded Data" display raw data retention levels, typically ranging from 70% to 90%. It demonstrates how, when taking co-coverage and joint feature availability into account, datasets with seemingly adequate data overall (such as "India") can have significantly less actual usable data. Furthermore, this filtering process's missingness threshold is a tunable hyperparameter, allowing users to modify sensitivity in accordance with particular requirements. This introduces an additional layer of interactivity, allowing balancing data volume against quality depending on applications. Overall, Figure 5 supports design goals G2 and G3 by quantifying missingness impact and guiding feature selection through clearer visibility into joint data integrity.

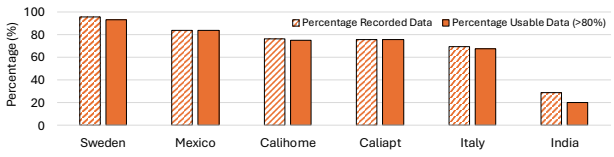

Fig. 5: Baseline Missingness Assessment Across Datasets

D. Scalablility for Efficient Data Processing

Figure 6 shows how our method significantly reduces the amount of data processed while maintaining analysis integrity, with the datasets grouped and standardized. The green bars show the actual rows analyzed after applying missingness and co-coverage filtering, compared to the grey bars, which represent the total dataset size. The reduction is more noticable in the larger datasets, showing our approach can scale effectively by minimizing unnecessary computations. The red line shows the duration of data collection, confirming that our approach works for both short-term and long-term datasets. Additionally, the figure shows that due to the non-uniform data collection frequencies, some datasets with shorter collection periods (e.g., "Calihome") still contain more rows than longer-term datasets (e.g., "India"). Our method bridges this gap by standardizing the data through resampling and coverage-based filtering, ensuring fair and consistent data quality assessment regardless of the original sampling rates. We compare the total number of rows per dataset before and after applying the

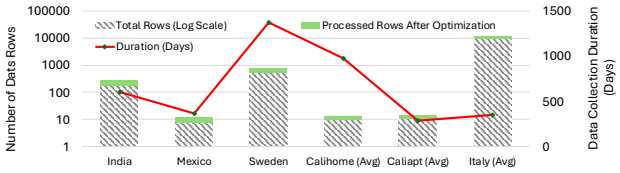

Fig. 6: Data Reduction across Datasets via our Framework

filtering process in Figure 6, showing that our method reduces the number of rows processed by a substantial margin while retaining high quality data segments. For more incomplete datasets (India, Mexico, Sweden, Italy), our method kept between 2% and 12% of rows; for more complete datasets (Calihome, Caliapt), it kept between 32% and 40%. With an average mean data reduction of 21%, the framework minimizes computational overhead and lets users concentrate on the most reliable segments. These results demonstrate that the framework is not only scalable but also effective in isolating high-quality multivariate time-series data from noisy and sparse collections . This figure demonstrates how the proposed approach takes into account G5 and G6, guaranteeing effective communication and scalable summarization in sizable, diverse datasets.

The results show our methods effectively meet the design requirements. Missing patterns are shown by temporal coverage heatmaps and period-specific missingness visualizations (G1), and its impact is measured by computed metrics (G2). The co-coverage matrix and parallel coordinates plot assess feature availability (G3) and assist in feature selection by highlighting strong co-coverage (G4). Data summarization is supported by multi-scale visualizations (G5), and scalability and interactivity are supported by the automated framework (G6). Although the parallel coordinates plot provides insightful information, its complexity indicates high-dimensional data requires better interactivity. Overall, the method offers a reliable way to assess the quality of data in multivariate time series.

V. Conclusion and Future Work

This article outlines a system designed to estimate the quality of data within multivariate time-series datasets, explicitly focusing on missingness, co-coverage, and scalability. Our framework ensures high-quality and consistent subsets for analysis through comprehensive metrics in conjunction with a heuristic model to systematically extract dependable data segments. Such interactions, together with a deep understanding of complex missingness, provide value with visualizations like temporal coverage heat maps and parallel coordinate plots, which lift the quality of data and help users make decisions on features and their usefulness for the intended purpose. In addition, the framework promotes scalability and handles the multi-scale data summarization. It addresses the tactics of overwhelming, high-dimensional datasets, which are common places for large-scale IoT environments with many limited-resource edge devices. By integrating visualizations with metrics, we balance automation with human involvement, helping experts to identify patterns and validate data subsets more reliably.

Although our framework is evidently effective, parallel coordinate plots' complexity remains a significant obstacle, particularly in high-dimensional settings. Furthermore, there needs to be a major improvement made to the current level of filtering and interaction. While future work will focus on addressing these gaps, some form of human analysis will still be required to validate the data and interpret the results. Our primary focus is on enhancing the interactivity of parallel coordinate plots by thoughtful incorporation of dynamic filtering, zooming, and clustering. We will also investigate more sophisticated summarization techniques to improve scalability and performance. To gain some transparency we will be adding tools for diagnostics on missing data and implementing automatic imputation based on co-occurrence patterns for more dependable data preprocessing. In addition, we are committed to adopting AI-based techniques, including machine learning models for intelligent missing data imputation, early prediction of data quality issues, and AI algorithms for automated feature selection and anomaly detection to streamline the subsequent phases of the analysis.

Acknowledgement

We thank the anonymous reviewers for their suggestions and feedback. This research was in part supported by US NSF Grants: SHF-2210744, AMPS-2229073.

References

[1] S. J. Fernstad and J. J. Westberg, "To explore what isn't there—glyph-based visualization for analysis of missing values," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 10, pp. 3513–3529, 2021.

[2] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 3, pp. 1249–1268, 2016.

[3] T. Gschwandtner and O. Erhart, "Know your enemy: Identifying quality problems of time series data," in *IEEE PacificVis*, 2018.

[4] S. Alsufyani, M. Forshaw, and S. J. Fernstad, "Visualization of missing data: a state-of-the-art survey," *arXiv:2410.03712*, 2024.

[5] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Proc. Int. Conf. AVI*, 2012.

[6] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer, "Visplause: Visual data quality assessment of many time series using plausibility checks," *IEEE Trans. Vis. Comput. Graph.*, 2016.

[7] M. Templ, A. Alfons, and P. Filzmoser, "Exploring incomplete data using visualization techniques," *Adv. Data Anal. Classif.*, vol. 6, no. 1, pp. 29–47, 2012.

[8] M. Theus, H. Hofmann, B. Siegl, and A. R. Unwin, "Manet: Extensions to interactive statistical graphics for missing values," in *New Techniques and Technologies for Statistics II*, 1997.

[9] N. J. Tierney and D. H. Cook, "Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations," *arXiv:1809.02264*, 2018.

[10] E. Jiménez and R. Macías, "Graphical tools for visualization of missing data in large longitudinal phenomena," in *Computer Graphics Forum*, vol. 41, pp. 438–452, Wiley Online Library, 2022.

[11] J. Honaker, G. King, and M. Blackwell, "Amelia ii: A program for missing data," *Journal of statistical software*, vol. 45, pp. 1–47, 2011.

[12] S. J. Fernstad, "To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization," *Information Visualization*, vol. 18, no. 2, pp. 230–250, 2019.

[13] S. Alsufyani, M. Forshaw, S. Del Din, A. Yarnall, L. Rochester, and S. J. Fernstad, "Multi-level visualization for exploration of structures in missing data," *CGVC. The Eurographics Association*, 2024.