Statistical Determinants of Optimal Minimum Segment Length for Time-Series Data Volume Reduction

Aleix Boixader Coma[1]✉[0009−0002−8519−589X], Rizkallah Touma[1][0000−0003−0543−1275], and Karina Gibert[2][0000−0002−8542−3509]

[1] Fundació i2CAT, Barcelona, Spain
{aleix.boixader,rizkallah.touma}@i2cat.net
[2] Universitat Politècnica de Catalunya, Barcelona, Spain
karina.gibert@upc.edu

Abstract. Data volume reduction is essential for managing the massive volume of high-frequency time-series data generated by modern sensor networks. However, the efficacy of such techniques often hinges on manual hyperparameter tuning, limiting their practical deployment. This study addresses this limitation by identifying the statistical determinants of the Optimal Minimum Segment Length (OMSL) hyperparameter of the Trend-Aware Piecewise Linear Approximation (TA-PLA) algorithm. Our approach includes a comprehensive empirical analysis of 31 statistical metrics that measure different aspects of time series data, such as stationarity, signal-to-noise ratio and spectral complexity. Using two heterogeneous datasets of physiological monitoring and datacenter performance time series, our results identify Allan Variance and Frequency-based Signal-to-Noise Ratio as primary, non-redundant indicators of the optimal balance between data reduction and reconstruction fidelity. By establishing this link between intrinsic time series characteristics and algorithm parameters, we provide a data-driven framework that eliminates the need for manual fine-tuning, significantly enhancing the usability and generalization of TA-PLA for autonomous data volume reduction.

Keywords: Time-series · Data Volume Reduction · TA-PLA · Hyperparameter Optimization · Statistical Characterization

1 Introduction

As the majority of enterprise data is now generated at the edge [6], time-series analysis has become fundamental for transitioning from raw observation to predictive modeling. However, time series data is frequently characterized by high redundancy and stochastic noise. Processing this low-value data incurs substantial hidden costs, as the transmission, storage, and utilization of unrefined time series in machine learning represent a significant source of computational and energy inefficiency. To mitigate this burden, the development of effective data reduction mechanisms is essential. The distinction between data reduction

strategies primarily involves the tradeoff between bit-for-bit reconstruction and compression efficiency, with research increasingly pivoting from entropy-limited lossless methods toward lossy frameworks for massive datasets [8]. By selectively discarding non-essential information, these methodologies alleviate the systemic strain on storage and processing architectures, making them indispensable to handle the scale of modern data [4,7].

In our previous work [1], we introduced the Trend-Aware Piecewise Linear Approximation algorithm, a two-stage segmentation framework designed for high-fidelity data volume reduction. The algorithm regulates the granularity of the segmentation by enforcing a Minimum Segment Length (MSL). This parameter acts as a lower temporal bound that prevents the model from reacting to high-frequency noise or stochastic oscillations and effectively merges non-informative local fluctuations into stable, trend-aligned linear segments while preserving critical structural features. This targeted reduction allows for significant data volume gains with minimal impact on signal integrity. Previous results demonstrated a reconstruction fidelity of 96% and only a marginal -0.0063 reduction in model R^2 scores, confirming that the systematic exclusion of these high-frequency components does not compromise downstream analytical tasks.

In this paper, we address the challenge of optimizing this parameter by characterizing the relationship between intrinsic time series properties and the Optimal Minimum Segment Length. Rather than relying on trial-and-error, we conduct a systematic empirical analysis to identify how standard, widely available statistical metrics correlate with the OMSL that maximizes reduction while maintaining a high correlation between original and reconstructed time series, thereby eliminating the need for expert calibration.

To perform this analysis, we utilized two independent time-series datasets with over 10 million observations. Our empirical evaluation reveals that the OMSL is highly sensitive to intrinsic signal properties. Specifically, we identified Stationarity and SNR, represented by log-transformed Allan Variance and Frequency-based SNR respectively, as primary descriptors of the optimal balance between data reduction and reconstruction fidelity. These results establish the groundwork for a more automated, data-driven application of the TA-PLA algorithm across diverse and unknown signal profiles.

2 Related Work

To manage the massive influx of time-series data produced by IoT sensor networks, specialized reduction methodologies are required. Chiarot et al. propose a taxonomy for time-series compression, evaluating algorithms based on their ability to maintain accuracy within the IoT ecosystem [3]. De Oliveira et al. identify the specific limitations of existing time-series solutions when deployed on resource-constrained devices, noting a growing trend toward using machine learning to handle these signals [11]. Pioli et al. highlight that applying reduction to heterogeneous sensor data is essential to avoid system bottlenecks and network congestion [12]. Furthermore, Serhii et al. propose distributed systems

that utilize neural networks to optimize the transmission efficiency of IoT data streams [16], while Pioli et al. examine the intersection of AI and dimensionality reduction for managing massive sensor volumes [13].

Beyond traditional methods, recent breakthroughs in generative modeling and non-linear transform coding have introduced a "neural compression" paradigm that replaces traditional algorithms with learned functions [2,17] or Large Language Models (LLMs) [9]. These methods are characterized by a computationally-heavy nature, often requiring significant GPU resources and high energy consumption for both training and real-time inference. Podgorelec et al. provide a comprehensive overview of these general trends, highlighting how cloud and edge computing are driving the shift toward feature-based compression and the integration of machine learning into universal data reduction frameworks [14].

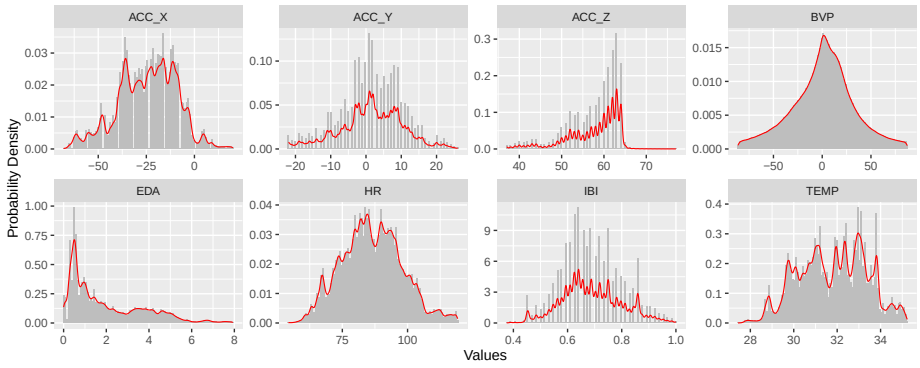
In contrast to the aforementioned works that rely on complex AI model architectures or require extensive manual calibration to achieve optimal results, this paper focuses on understanding the hyperparameter fine-tuning of our previously established TA-PLA algorithm [1]. We leverage standard statistical dataset metrics that are computationally inexpensive to determine the OMSL. By grounding the MSL optimization in accessible statistical features, we significantly enhance the usability and generalization of the approach across diverse datasets. This exploration establishes the groundwork to prevent suboptimal parametrization, making high-fidelity data reduction accessible for resource-constrained IoT environments without the need for manual heuristic tuning.

3 Datasets

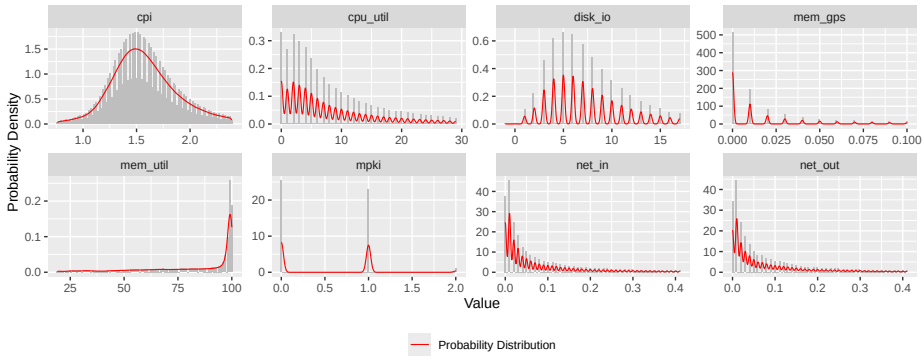
To evaluate the versatility and robustness of our approach, we included two contrasting, publicly available datasets: a Medical Monitoring Dataset and a Datacenter Performance Dataset. This selection allows us to validate the results of our empirical analysis against highly heterogeneous time series, ranging from the semi-periodic, smooth patterns of medical data to the bursty, stochastic trends characteristic of cloud computing environments.

- **Medical Monitoring Dataset.** This dataset was collected as part of a study to show the relationship between incorrect work place ergonomics and cognitive difficulties [10]. It consists of 8 physiological signals collected through the Embrace E4 wearable bracelet [3] with a total of 1,939,599 observations spread across 100 individuals, resulting in 487 time series. Fig. 1a illustrates the probability density distributions for these physiological variables, highlighting the non-Gaussian nature of signals like *EDA* and *TEMP*.
- **Datacenter Performance Dataset.** Published by Alibaba as an initiative to improve resource efficiency in datacenters[4], this dataset includes traces of 8 performance metrics collected in one of their cloud clusters for a duration of eight days. For our analysis, we sample 10,000 containers from the dataset,

[3] https://www.empatica.com/research/e4/
[4] https://github.com/alibaba/clusterdata



(a) Medical Monitoring Dataset



(b) Datacenter Performance Dataset

Fig. 1: Probability distributions of the datasets' variables.

obtaining a total of 9,475,892 observations divided into 77,766 time series. Fig. 1b presents the probability density distributions of the metrics, showing the heavy tail distributions of variables such as *mem_ util* and *net_ in*.

4 OMSL Formulation and Selection

4.1 OMSL Formulation

Our previous work [1] showed that for some time series, a conservative MSL selection produced a quasi-identical reconstruction but a moderate volume reduction percentage. This indicates that the MSL parameter could be selected to better fit each time series to optimize reduction percentage while maintaining reconstruction fidelity. Due to this, we hypothesize that there is an optimal value of MSL, called Optimal Minimum Segment Length, that maximizes the performance of the TA-PLA algorithm considering the following metrics:

- **Reduction Ratio.** Calculated as $(1 - 4R/T) \times 100$, where T is the original time series length, R is the number of segments of the reduced representation and 4 is the number of coefficients required to reconstruct each segment.
- **Correlation.** We use Pearson's absolute correlation to measure the linear similarity between the reconstructed and original signal, taking values in the range $[0.0, 1.0]$. Values closer to 1.0 indicate higher reconstruction fidelity.

We define the OMSL of a time series as the msl within a set of possible values $\mathcal{M}$ that maximizes the reduction percentage while maintaining high reconstruction fidelity, which we define as a correlation above 0.9 between the reconstructed and original time series. We formalize this heuristic in Eq. 1, where we also include a fallback in cases where no msl is found to maintain the high correlation threshold. In those cases, we select the smallest msl value that achieves a positive reduction percentage, meaning that we choose the highest correlation possible in exchange of a moderate reduction ratio.

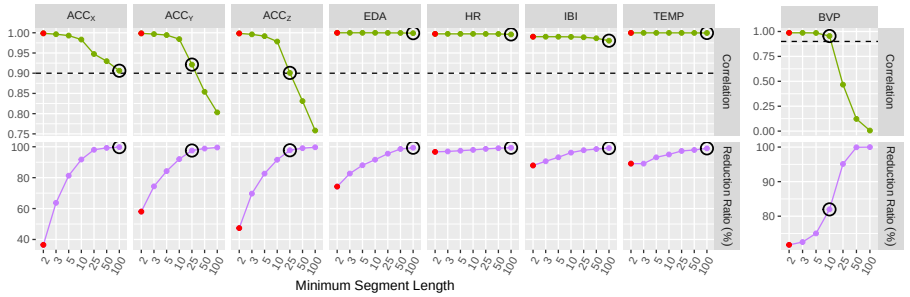
$$OMSL = \begin{cases} \arg\max_{msl \in \mathcal{M}} \{\text{Red}_{msl} \mid \text{Corr}_{msl} > 0.9 \wedge \text{Red}_{msl} > 0\} & \text{if} \\ \qquad\qquad \exists\, msl : (\text{Corr}_{msl} > 0.9 \wedge \text{Red}_{msl} > 0) & \\ \min\{msl \in \mathcal{M} \mid \text{Red}_{msl} > 0\} & \text{otherwise} \end{cases} \tag{1}$$

4.2 OMSL Selection

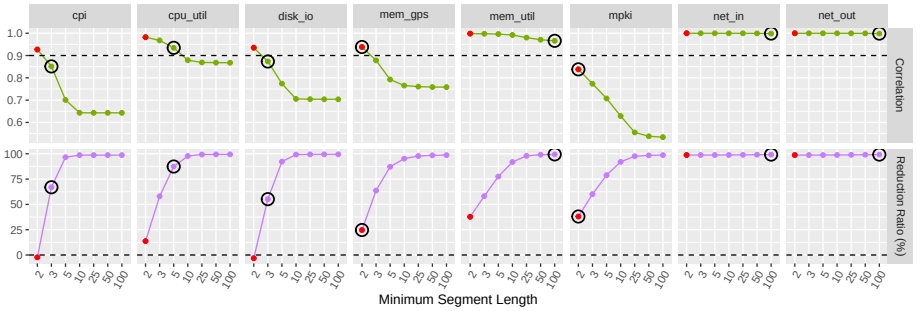
Following Eq. 1, we conducted an analysis to determine the OMSL of the variables presented in Section 3 using a search space of $\mathcal{M} = \{2, 3, 5, 10, 25, 50, 100\}$. Results were first computed for individual time series, i.e. per *subject* in the medical dataset and per *container* in the datacenter dataset, and subsequently averaged to characterize the overall performance for each variable.

Fig. 2 illustrates the evolution of correlation and reduction rate across different MSL values for each variable. The results indicate a clear bimodal behavior among variables: some maintain a stable correlation above 0.9 even as MSL increases, while others exhibit a sharp performance degradation beyond a specific threshold. This confirms that a universal OMSL is insufficient for diverse signal profiles. Table 1 quantifies these findings, providing the specific OMSL, reduction rate, and achieved correlation for each variable.

Individual analysis reveals specific edge cases. For instance, *mpki* fails to reach the 0.9 correlation threshold even at the minimum $MSL = 2$. This is attributed to its binomial distribution (see Fig. 1b), where the signal oscillates between discrete values; any linear segmentation inherently introduces reconstruction noise, though a correlation of 0.84 remains achievable. Similarly, *disk_io* and *cpi* require an $MSL = 3$ to achieve a positive reduction ratio, sacrificing slight fidelity for practical data volume gains. In contrast, *BVP* exhibits high compressibility but is sensitive to MSL increases; because it oscillates around zero, a high MSL causes our algorithm to treat these fluctuations as noise, leading to an OMSL of 10 to balance its 82% reduction with a 0.95 correlation.



(a) Medical Monitoring Dataset



(b) Datacenter Performance Dataset

Fig. 2: Evolution of reduction ratio and correlation with increasing MSL per variable. Dashed lines indicate the 0.9 correlation and 0 reduction thresholds. OMSL is marked with a black circle, while a red dot marks the global fixed MSL from previous work [1]. *BVP* is separated due to scale differences.

Table 1: Quantitative results for all variables using the OMSL found empirically.

	Variable	OMSL	Reduction	Correlation		Variable	OMSL	Reduction	Correlation
Medical Dataset	ACC_X	100	99.77%	0.91	**Datacenter Dataset**	cpi	3	67.03%	0.85
	ACC_Y	25	97.58%	0.92		cpu_util	5	87.61%	0.93
	ACC_Z	25	97.74%	0.90		disk_io	3	55.14%	0.87
	BVP	10	82.02%	0.95		mem_gps	2	24.86%	0.94
	EDA	100	99.26%	0.99		mem_util	100	99.52%	0.97
	HR	100	99.63%	0.99		mpki	2	38.09%	0.84
	IBI	100	99.14%	0.98		net_in	100	99.32%	0.99
	TEMP	100	98.72%	0.99		net_out	100	99.32%	0.99

These empirical results underscore the high variance in optimal parametrization across different time series. This heterogeneity motivates a deeper investigation into how intrinsic signal characteristics dictate the OMSL. Specifically, the following sections explore which statistical features of time-series determine the OMSL.

5 Statistical Characterization Metrics of Time Series

To understand the drivers behind OMSL variance, we evaluate a suite of statistical metrics designed to provide a comprehensive characterization of a time series' temporal changes. Given that the TA-PLA algorithm's performance depends on its ability to identify meaningful trend changes amid local fluctuations, we select a total of 31 metrics categorized into four primary domains. The main metrics are formally defined in Table 2. The names provided in parentheses throughout this section correspond to the shorthand identifiers used to refer to these metrics in the subsequent analysis and results.

1. **Stationarity:** These metrics assess the stability of the time series over time. Since our algorithm relies on monotonic directionality, understanding the degree of non-stationarity is crucial for determining how long a linear approximation (i.e. a segment) remains valid before a structural break occurs.
   - Local mean variation ($sd_local_mean_\{w\}$): Measures the standard deviation of means calculated over non-overlapping sliding windows of sizes $w \in \{10, 20, 50, 150, 300, 900\}$.
   - Allan Variance (avf_score) and its $\log_{10}$ (log_avf_score): Evaluates stability by comparing consecutive segment means. Determines the probability that a time series is stationary, unstable, non-stationary or trend-defined across different timescales.
   - Short-Term Variability ($sd1$) and its inverse ($inverse_sd1$): Quantifies high-frequency fluctuations (jitter).
   - Long-Term Variability ($sd2$) and its inverse ($inverse_sd2$): Captures the total dispersion and low-frequency components (trend).
   - Poincaré Unpredictability (sd_ratio) and its inverse ($inverse_sd_ratio$): It characterizes time series irregularity by calculating the $sd1/sd2$ ratio.
   - Hjorth's Activity ($hj_activity$) and its inverse ($inverse_hj_activity$): Interprets variance (σ^2) as the intensity of the data fluctuations.
   - Hjorth's Mobility ($hj_mobility$) and its inverse ($inverse_hj_mobility$): Represents the proportion of standard deviation attributed to the time series' rate of change
   - Hjorth's Complexity ($hj_complexity$): It compares the mobility of the time series' derivative to the time series itself to quantify how much the time series deviates from a pure sine wave.
2. **Signal-to-Noise Ratio (SNR):** These metrics quantify the relationship between underlying trends and stochastic interference of a time series. This is essential for evaluating the MSL's role as an implicit low-pass filter, as it identifies the threshold where the parameter begins to suppress informative signal components alongside noise.
   - Standard SNR (snr_s) its $\log_{10}$ (log_snr_s): It characterizes the time series strength relative to the total dispersion of the dataset.
   - Autocorrelation-based SNR (snr_td): It estimates the time series' persistence against uncorrelated white noise.
   - Frequency-based SNR (snr_f) and its $\log_{10}$ (log_snr_f): Compares the maximum spectral power against the mean power of the noise floor.

Table 2: Formal definitions of the statistical metrics used in our study, where: y is a time series of length T with individual observations y_t and first-order differential Δy_t. μ and σ represent the arithmetic mean and standard deviation of all the population, respectively. while $\bar{y}$ refers to the local arithmetic mean. $M = \lfloor T/w \rfloor$ is the number of segments for a window of size w. K is the Allan Variance sample count and ρ_1 is the 1-lag autocorrelation. P_{f_i} is the power at frequency bin i, f_{peak} is the frequency of maximum power, and F is the total number of bins. N denotes the number of entropy bins with associated probabilities p_i.

Statistical Metric	**Equation**
Local mean variation	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\bar{y}_m - \mu)^2}$
Allan variance	$\frac{1}{2(K-1)}\sum_{k=1}^{K-1}(\bar{y}_{k+1} - \bar{y}_k)^2$
Short-term variability	$\sqrt{\frac{1}{2}\mathrm{Var}(\Delta y_t)}$
Long-term variability	$\sqrt{2\mathrm{Var}(y_t) - \frac{1}{2}\mathrm{Var}(\Delta y_t)}$
Hjorth's Mobility	$\sqrt{\frac{\mathrm{Var}(\Delta y_t)}{\mathrm{Var}(y)}}$
Hjorth's Complexity	$\frac{\mathrm{Mobility}(\Delta y_t)}{\mathrm{Mobility}(y)}$
Autocorrelation (ρ_1)	$\frac{\sum_{t=2}^{T}(y_t-\mu)(y_{t-1}-\mu)}{\sum_{t=1}^{T}(y_t-\mu)^2}$
Standard SNR	$20\log_{10}\frac{\lvert\mu\rvert}{\sigma}$
Autocorrelation-based SNR	$10\log_{10}\left(\frac{\rho_1}{1-\rho_1}\right)$
Frequency-based SNR	$10\log_{10}\left(\frac{P_{f_{peak}}}{\frac{1}{F-1}\sum_{i=1,i\neq peak}^{F}P_i}\right)$
Relative Shannon Entropy	$\frac{-\sum_{i=1}^{N}p_i\log_2(p_i)}{\log_2(N)}$
Skewness	$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t-\mu}{\sigma}\right)^3$
Kurtosis	$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t-\mu}{\sigma}\right)^4$

3. **Spectral Distribution:** These metrics analyze a time series's frequency content and dominant periodicities to provide insight into its complexity. This allows us to characterize whether a time series is dominated by slow-moving trends or high-frequency oscillations that would be sensitive to a larger OMSL.
   - Autocorrelation (*hf_global_proxy*) and its inverse (*inverse_hf_global_proxy*): Measures the self-similarity of the time series at a 1-lag interval.
   - Relative Shannon Entropy (*entropy*): Quantifies the uncertainty of the time series's value distribution using a 100-bin histogram.
4. **General Descriptive Statistics:** This category includes fundamental metrics that provide a baseline for time series complexity and dispersion, capturing global characteristics that may not be fully described by specialized temporal or spectral analyses.
   - Standard deviation (*sd*): Quantifies the overall dispersion of the time series (σ).
   - Mean (*mean*): Represents the central tendency (μ) of the signal.

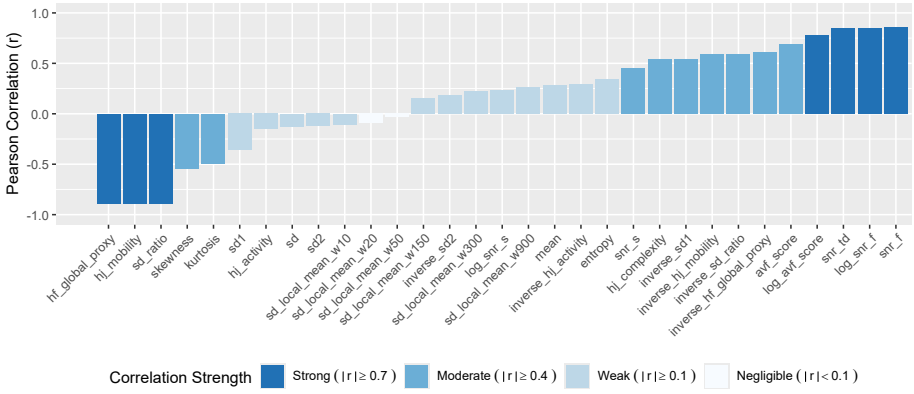


Fig. 3: Pearson's Correlation (r) between each statistical metric and $\text{OMSL}_{\log}$.

- Skewness (*skewness*): Quantifies the asymmetry of the value distribution relative to the mean.
- Kurtosis (*kurtosis*): Measures the "tailedness" of the distribution to identify the prevalence of outliers.

6 Statistical Determinants of OMSL

To identify which statistical metrics are the most determinant for the OMSL, we perform a Pearson correlation analysis between each metric and the OMSL across all 16 variables (Section 3). Because the OMSL values are drawn from the discrete set $\mathcal{M}$, we apply a $\log_{10}$ transformation to the target space, denoted hereafter as $\text{OMSL}_{\log}$.

Figure 3 illustrates these correlations, categorized according to the thresholds defined by Schober et al. [15]. Within this framework, a strong correlation ($|r| \geq 0.7$) indicates that a metric accounts for at least 50% of the variance in $\text{OMSL}_{\log}$. We identified 7 metrics meeting this criterion. Because *snr_f* and its logarithmic transformation *log_snr_f* are both strongly correlated to $\text{OMSL}_{\log}$, we retain only the base metric to avoid redundancy, finalizing a set of 6 significantly-correlated metrics.

Figure 4 illustrates the relationship between each significant statistical metric and the $\text{OMSL}_{\log}$ for each time-series variable, ordered by the absolute Pearson Correlation (r). Two variables, *mem_util* and *BVP*, consistently stand out as outliers, sitting far from the regression line. Regarding *BVP*, this divergence stems from the physiological nature of Blood Volume Pressure signals, which are defined by a normal distribution (Fig. 1a) with four distinct periodic phases that create frequent monotonic trend changes at a rate of two to four per second[5]. Due to this, increasing the MSL has a disproportionate effect on BVP; once the MSL exceeds the natural heart periodicity, the algorithm oversimplifies the signal and causes a significant loss of information. Consequently, the OMSL for

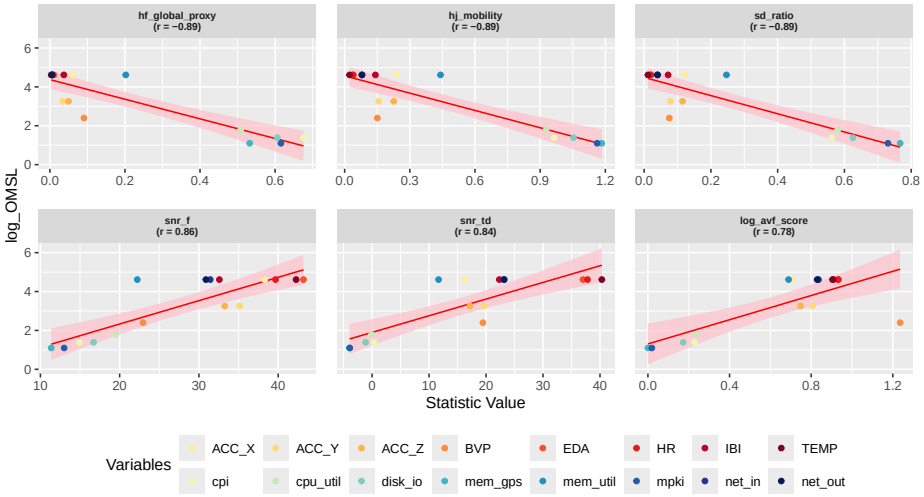


Fig. 4: Correlation of the significantly-correlated statistical metric with the OMSL_{log}. Each point represents a time series variable. The red line and shaded background correspond to the linear regression and its standard error.

BVP remains overly conservative compared to the rest of the dataset, resulting in it not entirely fitting the regression line of otherwise similar variables.

On the other hand, the underlying behavior of *mem_ util* is due to the specific role of memory as a datacenter bottleneck, where the pursue of infrastructure efficiency pushes servers toward maximum capacity. This leads to time series that are largely stationary at 100% with a long sparse tail extending over lower values (Fig. 1b). This heavily-skewed distribution allows us to significantly increase the MSL without a corresponding drop in reconstruction fidelity, again resulting in the OMSL of *mem_ util* standing out against the regression line.

Additionally, the top three statistical metrics most correlated with OMSL exhibit a striking similarity in their distributions, with data points clearly clustered into two distinct groups. Since these groupings correspond to the same variables across all three metrics, it suggests they are measuring the same underlying phenomenon. To address this likely redundancy, we performed a cross-correlation analysis between the 6 top-performing determinants. We applied a systematic pruning process by calculating a correlation matrix for all candidate metrics with a cutoff threshold of 0.8. For any pair exceeding this limit, we calculated the mean absolute correlation of each of them against all others in the set. The metric with the higher average redundancy was removed, and the process was repeated iteratively until no pairs remained above the threshold.

Figure 5 visualizes this metric-selection process, where the size of the circles represents the absolute magnitude of the correlation coefficient $|r|$ and color represents the direction and strength of the relationship. The hierarchical clustering identifies a dense redundancy group in the bottom-left; we can see *log_ avf_ score*

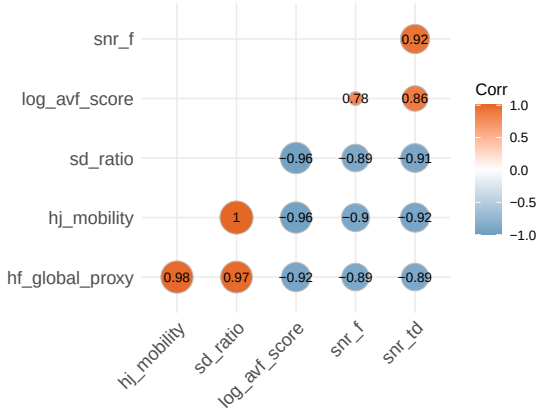


Fig. 5: Correlation matrix between the most significant statistical metrics.

belongs to this cluster because it shares near-perfect correlations ($|r| = 0.96$) at its intersections with *hj_mobility* and *sd_ratio*. However, it was selected as the group's representative because it is the most mathematically distinct; this is visually evident as its circle sizes shrink (e.g., to 0.78) when compared to other metrics, whereas variables like *sd_ratio* maintain large, dark circles across almost the entire set. Similarly, *snr_f* was retained over *snr_td* because it displays lower average absolute correlation. Overall, this analysis indicates that the log-transformed Allan Variance (*log_avf_score*) and Frequency-based Signal-to-Noise Ratio (*snr_f*) are the most suitable, non-redundant OMSL determinants, having the lowest correlation identified in Fig. 5 by the smallest circle size.

7 Conclusions

This research successfully establishes a structured, data-driven framework that lays the foundation for determining the hyperparameter selection of the TA-PLA algorithm. Our empirical evaluation reveals that the Optimal Minimum Segment Length is highly dependent on intrinsic signal properties, which are determined by stationarity, represented by log-transformed Allan Variance, and Signal-to-Noise Ratio, in particular Frequency-based SNR. Future work will focus on integrating an adaptive tuning module into the TA-PLA framework to calculate these metrics, enabling autonomous OMSL selection across heterogeneous time-series profiles. Additionally, we consider extending the analysis to different time series domains, such as financial, industrial or environmental to evaluate the generalization of the determinants identified. Furthermore, a sensitivity analysis will be conducted to determine how sensistive are the conclusions to a modification of the correlation threshold.

Disclosure of Interests. Karina Gibert is a Program Committee member of the XIII Symposium on Theory and Applications of Data Mining (TAMIDA 2026). The authors have no other competing interests to declare that are relevant to this article.

References

1. Boixader Coma, Aleix and Touma, Rizkallah and Gibert, Karina: Reducing storage volume of sensor data towards sustainable data science modeling. In: Artificial intelligence research and development: proceedings of the 27th International Conference of the Catalan Association for Artificial Intelligence. pp. 32–41. IOS Press (2025)
2. Chen, J., Fang, Y., Khisti, A., Özgür, A., Shlezinger, N.: Information compression in the ai era: Recent advances and future challenges. IEEE Journal on Selected Areas in Communications (2025)
3. Chiarot, G., Silvestri, C.: Time series compression survey. ACM Computing Surveys **55**(10), 1–32 (2023)
4. Di, S., Liu, J., Zhao, K., Liang, X., Underwood, R., Zhang, Z., Shah, M., Huang, Y., Huang, J., Yu, X., et al.: A survey on error-bounded lossy compression for scientific datasets. ACM computing surveys **57**(11), 1–38 (2025)
5. Elgendi, M.: On the analysis of fingertip photoplethysmogram signals. Current cardiology reviews **8**(1), 14–25 (2012)
6. Forbes Technology Council: 2025 it infrastructure trends: The edge computing, hci and ai boom. Forbes (2024), https://www.forbes.com/councils/forbestechcouncil/2024/12/12/2025-it-infrastructure-trends-the-edge-computing-hci-and-ai-boom/
7. Hosseini, M.: A survey of data compression algorithms and their applications. arXiv preprint arXiv:2506.10000 (2025)
8. Jayasankar, U., Thirumal, V., Ponnurangam, D.: A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. Journal of King Saud University-Computer and Information Sciences **33**(2), 119–140 (2021)
9. Li, Z., Huang, C., Wang, X., Hu, H., Wyeth, C., Bu, D., Yu, Q., Gao, W., Liu, X., Li, M.: Lossless data compression by large models. Nature Machine Intelligence **7**(5), 794–799 (2025)
10. Markova, V., Ganchev, T., Filkova, S., Markov, M.: Mmd-msd: A multimodal multisensory dataset in support of research and technology development for musculoskeletal disorders. Algorithms **17** (5 2024)
11. de Oliveira, M.A., da Rocha, A.M., Puntel, F.E., Cavalheiro, G.G.H.: Time series compression for iot: A systematic literature review. Wireless Communications and Mobile Computing **2023**(1), 5025255 (2023)
12. Pioli, L., Dorneles, C.F., de Macedo, D.D., Dantas, M.A.: An overview of data reduction solutions at the edge of iot systems: a systematic mapping of the literature. Computing **104**(8), 1867–1889 (2022)
13. Pioli, L., de Macedo, D.D., Costa, D.G., Dantas, M.A.: Intelligent edge-powered data reduction: A systematic literature review. ACM Computing Surveys **56**(9), 1–39 (2024)
14. Podgorelec, D., Strnad, D., Kolingerová, I., Žalik, B.: State-of-the-art trends in data compression: Compromise case study. Entropy **26**(12), 1032 (2024)
15. Schober, P., Boer, C., Schwarte, L.A.: Correlation coefficients: appropriate use and interpretation. Anesthesia & analgesia **126**(5), 1763–1768 (2018)
16. Serhii, U., Vasyl, K.: Optimizing data transmission in iot networks through enhanced compression and edge computing techniques. In: 2023 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo). pp. 76–79. IEEE (2023)
17. Yang, Y., Mandt, S., Theis, L.: An introduction to neural data compression. Foundations and Trends in Computer Graphics and Vision **15**(2), 113–200 (2023)