MCAD: Multivariate Correlation Anomaly Data Generator

Christos C. Papadopoulos
Aalborg University, NKUA,
and Athena R.C.
Aalborg, Denmark
ccpa@cs.aau.dk

Alkis Simitsis
Athena R.C.
Athens, Greece
alkis@athenarc.gr

Torben Bach Pedersen
Aalborg University
Aalborg, Denmark

ABSTRACT

Multivariate time series are ubiquitous in large scale data analytics, yet most anomaly detection benchmarks ignore the multi-channel characteristics of occurring anomalies. This demonstration will showcase the Multivariate Correlation Anomaly Data (MCAD) generator, a synthetic data generator that produces data that contain anomalies which span multiple channels in a multivariate series. MCAD is highly customizable and can produce a wide range of typical correlation patterns within the generated series. At the first data generation step, MCAD generates multivariate events based on different event templates by injecting these templates into different channels in the multivariate series at the same time slice. In the second data generation step, MCAD injects unique events that encompass multiple channels and do not occur as part of the normal event pattern in the first step, corresponding to anomalous behaviour. The demonstration will focus on guiding the audience through the data generation process using MCAD-VIZ, MCAD's visualization module. During the first part of the demonstration, MCAD-VIZ will be used to produce various datasets that contain different, non-recurring anomalous events. In the second part, the audience will see how natural approaches to the anomaly detection process produce erratic results through demonstrating a profiling process which will evaluate the detection performance of a black-box anomaly detector on the generated dataset, showcasing MCAD's motivation and the need for new research initiatives in the field of multivariate time series anomaly detection.

PVLDB Reference Format:
Christos C. Papadopoulos, Alkis Simitsis, and Torben Bach Pedersen. MCAD: Multivariate Correlation Anomaly Data Generator. PVLDB, 19(12): 4638 - 4641, 2026.
doi:10.14778/3827998.3828085

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Christosc96/MCAD.

1 INTRODUCTION

Multivariate time series (MTS) are increasingly becoming more and more prevalent in numerous fields such as industrial IoT, health, data center monitoring etc. (see [3, 6] and the references therein). This has sparked new interest within the scientific community, related to the analysis, visualisation, storage and processing of data that result from multiple sources and constitute a multivariate series. One of the most important goals of time series analysis is the identification of anomalous parts of the time series, which has been the subject of extensive study in both industry and academia alike [5]. Naturally, time series anomaly detection is also of interest in the multivariate setting. In this setting, anomalies can manifest in different channels of the MTS, indicating different sources of anomalous behaviour. To the best of our knowledge, most of the work that has been done in multivariate anomaly detection has focused on anomalies in a single channel of the MTS which does not take into account the multivariate nature of the problem and can be efficiently solved by porting algorithms from the univariate case.

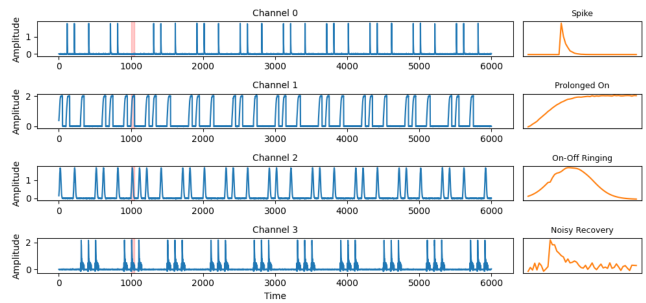

Figure 1: A representative example of a multivariate time series generated through MCAD. The time slice highlighted in red indicates an anomalous multivariate event.

A much more important problem is that of *correlation anomalies*, anomalies that manifest over specific subsets of channels in the MTS, and typically cannot be identified in any of these channels in isolation. As a concrete motivational example, consider the case of a wind turbine gearbox, equipped with multiple sensors that measure a multitude of parameters including axis vibrations, rotor speed, power output, and oil temperature. Recording these measurements over time constitutes a multivariate time series since each such measurement provides a distinct signal within the series.

While increases in temperature or increased vibrations may seem within normal thresholds, rises in temperature under otherwise normal load or vibration pattern shifts that occur with increased rotor speed may indicate an anomaly, such as early wear of gearbox parts. It is therefore important, not only to identify this anomaly, but also to isolate the channels that constitute the source of the anomaly, as remediation actions are different for each such source. To date, most anomaly detection datasets concerning time series, even multivariate ones, do not contain correlation anomalies or

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828085

even any at all, which can also explain why univariate counterparts of anomaly detection algorithms can also perform well in the multivariate setting [8].

We present the Multivariate Correlation Anomaly Data generator, MCAD. MCAD is a synthetic data generator that aims to drive research in problems concerning multivariate correlation anomalies, such as detection and root cause identification. MCAD builds upon and extends previous efforts in time series dataset generation [7] by allowing the user to generate MTS datasets of *any* arity. With MCAD, users can generate an MTS modeled using different event templates, as shown in Figure 2. MCAD will generate multivariate events that contain combinations of event templates, indicating inter-channel correlation. MCAD will then inject uniquely occurring combinations of events that are regarded as a *multivariate anomaly*. Our technical contributions are the following:

- **MCAD Generator.** MCAD is a novel data generator geared towards generating data that facilitate research concerning correlation anomalies in MTS. MCAD is developed as a Python module and is open source in order to foster research related to correlation anomalies in multivariate time series.
- **Extensive Customization.** MCAD allows for the injection of multivariate correlation anomalies of a specified channel arity, where only specified channels in the multivariate time series cause the anomalous event. The user can furthermore generate normal patterns, corresponding to predefined sets of events that indicate inter-channel correlation templates.
- **MCAD-VIZ.** MCAD features an interactive visualization component, MCAD-VIZ, to enable the inspection of various aspects of the generated data, normal patterns, and anomalies. During the use-case scenario presentation, MCAD-VIZ will be used to process specific channels of the generated MTS with a black-box oracle and see contrastive detection results that further accentuate the relevance of anomalous channel identification in correlation anomalies.

2 THE MCAD GENERATOR

MCAD is based on the generation of *multivariate events*, subsequences that indicate a combination of specific events that occur in different channels at the same time slice. These multivariate events can indicate, for example, increased vibrations due to an increase in rotor speed in wind turbines, which is a multivariate event involving two channels. MCAD generates *k-events*, multivariate events that occur in k different channels with k ranging from 1 to m, the total number of channels in the series. Multiple occurrences of the same k-event indicate a normal pattern, while uniquely occurring k-events are considered multivariate discords, since they are dissimilar to all other events in the series.

Next, we present the core components of MCAD. Figure 2 depicts the MCAD system architecture.

Univariate Event Templates In MCAD, the user can specify the univariate base and event templates for the generated data. Base templates define the basic oscillating signal of the series (alternatively, channels), which is the pattern that corresponds to any non-event part of the MTS. Event templates describe the form of events in the different channels of the MTS. These templates can be fully customised by the user to fit their needs and the relevant

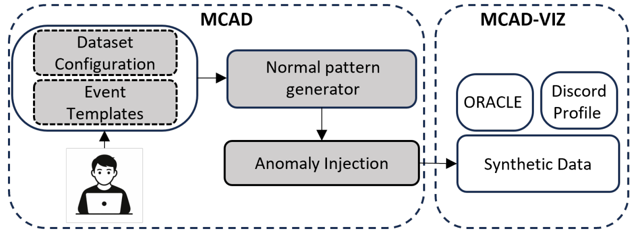

Figure 2: Architecture of MCAD.

application domain. For example, templates may indicate a heart pulse, on-off signals, noisy increase etc (see Fig.2), and the user can create and add fully custom templates to a generator instance through MCAD's Python API.

Normal Pattern Generator. The Normal Pattern Generator initiates the first step of the data generation process starts by the user specifying the channel-wise templates. The user then selects the discord's arity k, the number of channels that comprise the anomalous multivariate event that will be injected in the synthetic dataset generated by MCAD. The generator will then proceed to generate a *normal pattern*, a series of co-occurring channel-wise events that reflect normal functionality within the time series.

Anomaly Injection. The second component of MCAD is responsible for generating the anomalous k-event, which is an event with arity k. By definition, the normal pattern of events will contain all $(k-1)$-events generated within the multivariate series, so the k-event is unique in that it does not occur anywhere else in the series, representing a unique combination of events that deviate from the normal pattern generated so far.

MCAD-VIZ. MCAD allows the user to interact with generated data in order to better understand the generated dataset's properties regarding the generated normal pattern and anomalies injected in the data. Moreover, MCAD-VIZ offers direct anomaly detection capabilities on the generated data, by supporting *packaged black-box anomaly detectors* that can support a wide range of use-case scenarios.

3 DEMONSTRATABLE COMPONENTS

To highlight MCAD's synthetic data generation process and the core architectural and technical components of the proposed demonstration, we first begin by presenting the problem settings under which the demonstrated data generated using MCAD drive the need for novel research regarding the identification of multivariate anomalies and their root causes.

Multivariate Time Series. A multivariate time series $T \in \mathbb{R}^{m \times n}$ is a set of sequences of real-valued numbers with m channels where each sequence $T_j \in \{T_1, T_2, \dots, T_m\}$ represents a different variable. Each multivariate time series can be broken into *multivariate subsequences* of length w, $T_\ell \in \mathbb{R}^{m \times w}$. The total number of subsequences is $n - w + 1$ and ℓ indicates the starting position for a set of m subsequences belonging to different channels. MCAD generates and injects *sequence anomalies* in MTS that constitute *discords* [9]. In MTS, a discord is defined as follows.

Definition 1 (Multivariate discord). Given a multivariate time series $T \in \mathbb{R}^{m \times n}$, a window size w and an anomaly detection function f that takes T as input and produces an anomaly score, the multivariate discord of the series is the multivariate subsequence T_ℓ with the maximal anomaly score.

Multivariate discords typically manifest over a specific subset of channels, making the discord detection process more challenging, since irrelevant channels can obscure the true discord. That is, given a multivariate discord, there exists a set of channels $\mathcal{M}$ so that the following properties apply. For any set of channels $\mathcal{S} \subset \mathcal{M}$, $f(T_\ell^{(\mathcal{S})}) \ll f(T_\ell^{(\mathcal{M})})$. Similarly, for any superset $\mathcal{S}'$ s.t. $\mathcal{M} \subset \mathcal{S}'$, $f(T_\ell^{(\mathcal{S}')}) < f(T_\ell^{(\mathcal{M})})$. In other words, for any sets of channels that are a subset of the discord's channels, we observe a sharp decrease in magnitude. For channel supersets, the true magnitude of the multivariate discord is obscured, as irrelevant channels are expected to introduce noise in the anomaly detection process. Though MCAD is not specific to certain assumptions about the anomaly detection process, MCAD-VIZ integrates all anomaly detectors as black-box *Discord Oracles*.

Definition 2 (Discord Oracle). Given a MTS $T \in \mathbb{R}^{m \times n}$, a Discord Oracle is a black-box function $O : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ that takes as input T and produces an output anomaly score corresponding to the presence of a multivariate discord in T and its corresponding magnitude.

Generating normal patterns. As previously stated, the first step in MCAD's generation process invokes the Normal Pattern Generator, the component responsible for creating sets of events that constitute a normal pattern in the multivariate series. In order to do so, the Normal Pattern generation process takes as input the normality coefficient.

Definition 3 (Normality Coefficient). The normality coefficient of a MTS generated by MCAD is the required number, N, of occurrences of multivariate events, so that each event has at least N occurrences in the generated MTS.

The normality coefficient is a user defined parameter, and reflects the underlying normal model, with the aim of fitting the needs of multiple tasks. For example, a black-box oracle based on the Matrix Profile [9] may require a lower normality coefficient than an oracle based on a normal model identification [1, 2]. The normality coefficient dictates the minimum length of the generated time series that satisfies the coefficient's value. Specifically, given an event length w, a minimum gap between events g, a normality coefficient N and an anomaly arity of k

$$n \geq N \times (w + g) \times \left(\sum_{k_c=2}^{k-1} \binom{m}{k_c} + 1 \right) (1)$$

Discord Profiling. Once the dataset is generated, the MCAD-VIZ component allows the user to utilize discord oracles in order to inspect different channels of the MTS with regard to their anomaly scores. MCAD-VIZ does so by employing *Discord Profiling*, an anomaly analysis process based on outlierness profiles in subspace search visualization [4]. Given a set of channels of the MTS $\mathcal{M}$, discord profiling proceeds by instantiating a Discord Profile, which is a vector $P(\mathcal{M}) = \left(\mu_1(T^{(\mathcal{M})}), \mu_2(T^{(\mathcal{M})}), \ldots, \mu_m(T^{(\mathcal{M})})\right)$ which contains the mean anomaly scores of the discord oracle for all channel subsets and supersets of $\mathcal{M}$ with a specific arity, $\mathcal{S}_k(\mathcal{M}) = \{A \subseteq \{1, \ldots, m\} \ : \ |A| = k, \ (A \subseteq \mathcal{M} \ \text{ or } \ \mathcal{M} \subseteq A)\}$. Here, $\mu_k(T^{(\mathcal{M})}) = \frac{1}{|\mathcal{S}_k|} \sum_{S \in \mathcal{S}_k} O\big(T^{(\mathcal{M})}\big)$. Discord profiles provide information about the severity of a discord detected in specific channels of the MTS, and the relative severity of the discord manifesting over these specific channels when considering less or more channels than the given channel set. During the demonstration we will show that discord profiles for true multivariate discords peak at the discord's arity, while peaks of discord profiles of channels that do not contribute to the discord is considerably lower indicating normal behaviour.

4 DEMONSTRATION SCENARIOS

Figure 3 presents a screenshot of MCAD-VIZ depicting a MTS generated through MCAD. Through the configuration panel (Figure 3 (a)), the user can control the parameters of the data generation process, such as the total channels of the MTS and their templates, arity of the anomalous k-event and the normality coefficient. This allows MCAD to generate data fitting multiple use-case scenarios regarding the data generation process. For the demonstration, MCAD-VIZ comes with a pre-packaged Discord Oracle based on matrix profiling. In order to make the demonstration more intuitive, all event templates used will adhere to our wind turbine example use-case. We will start our demonstration with the following practical, and frequently met scenarios.

Global Multivariate Discords. Research and data revolving around Anomaly Detection in MTS often uses global discords, where the multivariate discord spans all MTS channels. In this scenario, the MTS is generated by setting $k = m$ in MCAD. Since the discord is not obscured by any irrelevant channels, and $O(T)$ will produce accurate results.

Obscured Multivariate Discords. MCAD's core technical contribution is best showcased when generating datasets with discords of arity $k < m$. As can be seen in Figure 3 (b), through MCAD-VIZ, the user can inspect the generated series and query for occurrences of selected multivariate events. In the Figure, a multivariate event spanning channels $\{0, 1, 2\}$ occurs only once, indicating that this event is a discord. This scenario aims to highlight limitations of existing approaches, as discovering such Multivariate Discords efficiently requires prior knowledge of the discord's arity to avoid exponential search of possible channel sets. Finally, the case of unknown arity requires the development of sophisticated channel selection policies and algorithms, which is the core motivational research aim of the demonstration.

Discord Profiling Results. In the second part of the demonstration, we will present the participants with a use-case scenario that involves using the generated MTS to conduct discord analysis with a given discord oracle. The objective of this part is to showcase the anomaly score degradation of the unique correlation discord when considering irrelevant channels. As shown in Figure 3 (c), participants can generate discord profiles for different channel sets of the generated MTS. As explained in Section 3, peaks in a discord profile indicate true multivariate discords, while channel supersets introduce noise to the real discord, lowering the anomaly score resulting from the discord oracle. Subsets and disjoint sets of the true discord's channels will appear regular, with little to no anomalous behaviour indicated by the oracle. In Figure 3, channels of arity 1 and 2 appear normal, while the consideration of the full MTS lowers the true discord's anomaly score.

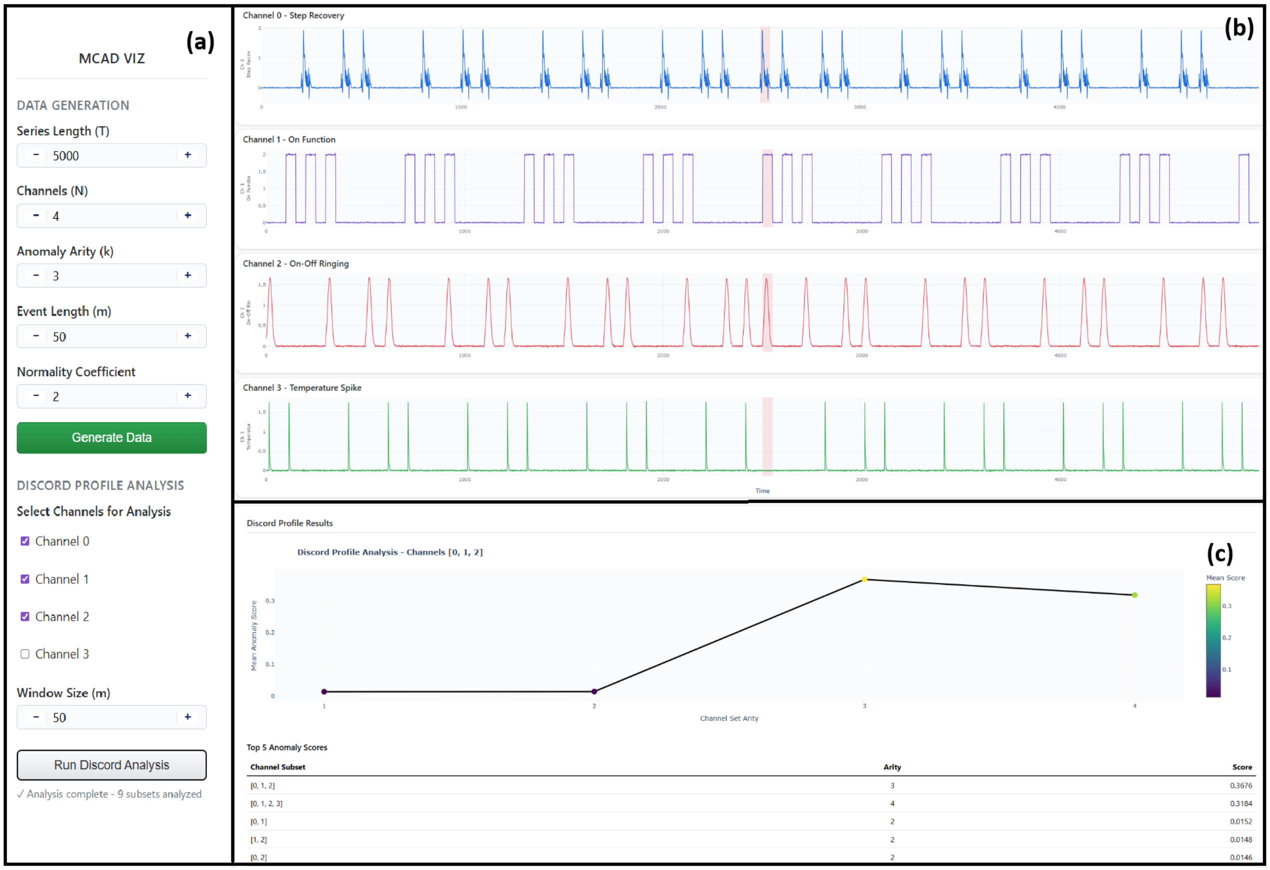

Figure 3: Overview of MCAD's data generation and Discord Analysis with MCAD-VIZ.

Incomplete Normal Patterns. As incomplete normal patterns can appear in applications involving streaming or noisy data, MCAD can also generate incomplete patterns that violate Equation 1, resulting in generated data that can also contain $(k-1)$-events that appear as anomalous. In this case, we demonstrate that Discord Profiling results in erratic findings, since there is no way to distinguish events that appear anomalous from true multivariate discords.

Off-script demonstration. Finally, interested participants will be able to generate their own datasets, investigate ad-hoc cases and fully explore the range of MCAD's capabilities, as a way to highlight MCAD's general applicability and rich configuration options.

5 CONCLUSIONS

MCAD is a data generator that produces Multivariate Time Series data targeted towards Multivariate Correlation Anomalies. MCAD supports custom event templates, and can generate Multivariate Discords that are obscured when considering the entire MTS and provides a data analytics interface that assists users with investigation of anomalous multivariate events, motivating the need for novel approaches in Multivariate Anomaly Detection.

Acknowledgments. This work has been partially supported by the H2020-MSCA-ITN-2020 DEDS project (GA 955895) and the EU Horizon Europe DataGEMS project (GA 101188416).

REFERENCES

[1] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal* 30, 6 (2021), 909–931.

[2] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. 2021. SAND: Streaming Subsequence Anomaly Detection. *Proc. VLDB Endow.* 14, 10 (2021), 1717–1729.

[3] Jens E d'Hondt, Haojun Li, Fan Yang, Odysseas Papapetrou, and John Paparrizos. 2025. A structured study of multivariate time-series distance measures. *Proceedings of the ACM on Management of Data* 3, 3 (2025), 1–29.

[4] Fabian Keller, Emmanuel Müller, Andreas Wixler, and Klemens Böhm. 2013. Flexible and adaptive subspace search for outlier analysis. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* 1381–1390.

[5] Qinghua Liu and John Paparrizos. 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems* 37 (2024), 108231–108261.

[6] John Paparrizos, Paul Boniol, Qinghua Liu, and Themis Palpanas. 2025. Advances in time-series anomaly detection: Algorithms, benchmarks, and evaluation measures. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 2.* 6151–6161.

[7] Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. [n.d.]. TimeEval: A Benchmarking Toolkit for Time Series Anomaly Detection Algorithms. 15, 12 ([n. d.]), 3678 – 3681. https://doi.org/10.14778/3554821.3554873

[8] Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. 2024. Anomaly detectors for multivariate time series: the proof of the pudding is in the eating. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW).*

[9] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM).*