Orange Research

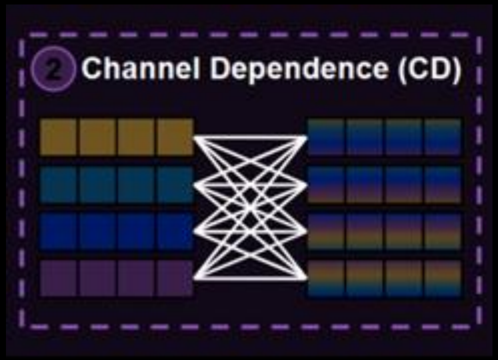
Anomalies in Multivariate Time Series Anomaly Detection Benchmarks Are Mostly Univariate

Marc PINET [1,2]
Julien CUMIN [1]
Samuel BERLEMONT [1]
Dominique VAUFREYDAZ [2]

[1] Orange Research

[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG

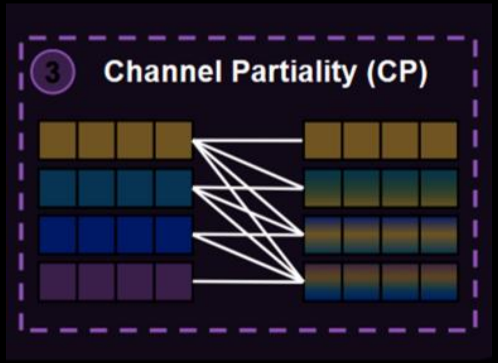
State-of-the-art of Multivariate Anomaly Detection: CI vs CD vs CP



Channel-dependent (CD)

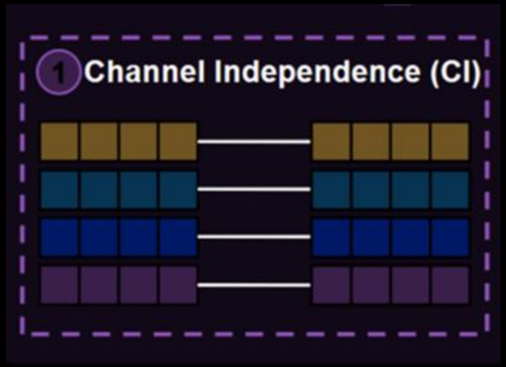
OmniAnomaly *(Su et al. 2019)*
MSCRED *(Zhang et al. 2019)*
USAD (Audibert et al. 2020)
TranAD *(Tuli et al. 2022)*
… and more

Original way of training TS models (before CI and CP)



Channel-partiality (CP)

Graph-based: GDN *(Deng & Hooi 2021)*, MTAD-GAT *(Zhao et al. 2020)*, …
MLP-mixers: PatchAD *(Zhong et al. 2025)*, CCM-TAD *(Murad & Yilmaz 2025)*, …
Cross-channel Attention: CATCH *(Wu et al. 2025)*, XCTFormer *(Zexer & Azencot 2026)*, …
Hybrid CI/CD: DBAD *(Sun et al. 2025)*
… and more



Channel-independent (CI)

CrossAD *(Li et al. 2025)*
DADA *(Shentu et al. 2025)*
DCdetector *(Yang et al. 2023)*
KAN-AD *(Zhou et al. 2025)*
CiTranGAN *(Chen et al. 2025)*
… and more

Emerging trend since PatchTST (Nie et al. 2023) in forecasting

Source: A Comprehensive Survey of Deep Learning for MTS Forecasting: A Channel Strategy Perspective, Qiu et al. (2025)

Evaluation using the same public benchmarks

GECCO (Water Quality Dataset) - *Moritz et al. 2018*
MSL (Mars Science Laboratory) - *Hundman et al. 2018*
SMAP (Soil Moisture Active Passive) - *Hundman et al. 2018*
PSM (Pooled Server Metrics) - *Abdulaal et al. 2021*
SMD (Server Machine Dataset) - *Su et al. 2019*
SWAN-SF (Space Weather ANalytics for Solar Flares) - *Angryk et al. 2020*
SWaT (Secure Water Treatment) - *Goh et al. 2017*
WADI (WAter DIstribution) - *Ahmed et al. 2017*

Dataset	C	Train size	Test size	Anomaly ratio (%)
GECCO	9	69 260	69 261	1.05
MSL	55	58 317	73 729	10.53
SMAP	25	140 825	444 035	12.87
PSM	25	132 481	87 841	27.76
SMD	38	708 405	708 420	4.16
SWAN-SF	38	60 000	60 000	32.60
SWaT	31	472 459	472 460	11.56
WADI	79	784 571	172 803	5.77

The problem

On these benchmarks, CI beats CD and CP, why?

Known suspicions were never quantified

- **Wu & Keogh 2021** **(Univariate scope only) Benchmarks are flawed**

- **Garg et al. 2022** **A univariate AE beats SOTA models**
- **Wenig et al. 2024** **Univariate detectors beat multivariate ones on real benchmarks**

Explanatory hypothesis

- Garg 2022, Wenig 2024 **If univariate (CI) models already perform well on these benchmarks…**

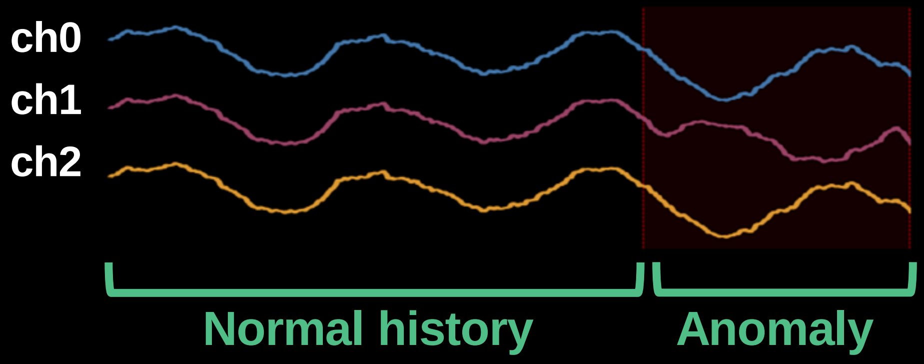
… do these benchmarks require cross-channel modeling at all?

Let's infer directly on the data rather than on detector performances…

Methodology
Characterize the dataset, not the detector!

For each labeled anomalous segment:

1. *Define normal history*
2. *Use history to characterize*
3. *Classify into 4 categories*



Characterizations with lag support

- *univariate*
  density of univariate timesteps
- *cross-channel structure*

Univariate Test
Univariate Ratio
Correlation Test

Classification

Table 2: Per-segment classification combining the two tests.

	Univariate ✓	Univariate ×
Cross-channel ✓	BOTH	CROSS-CHANNEL
Cross-channel ×	UNIVARIATE	UNDETECTED

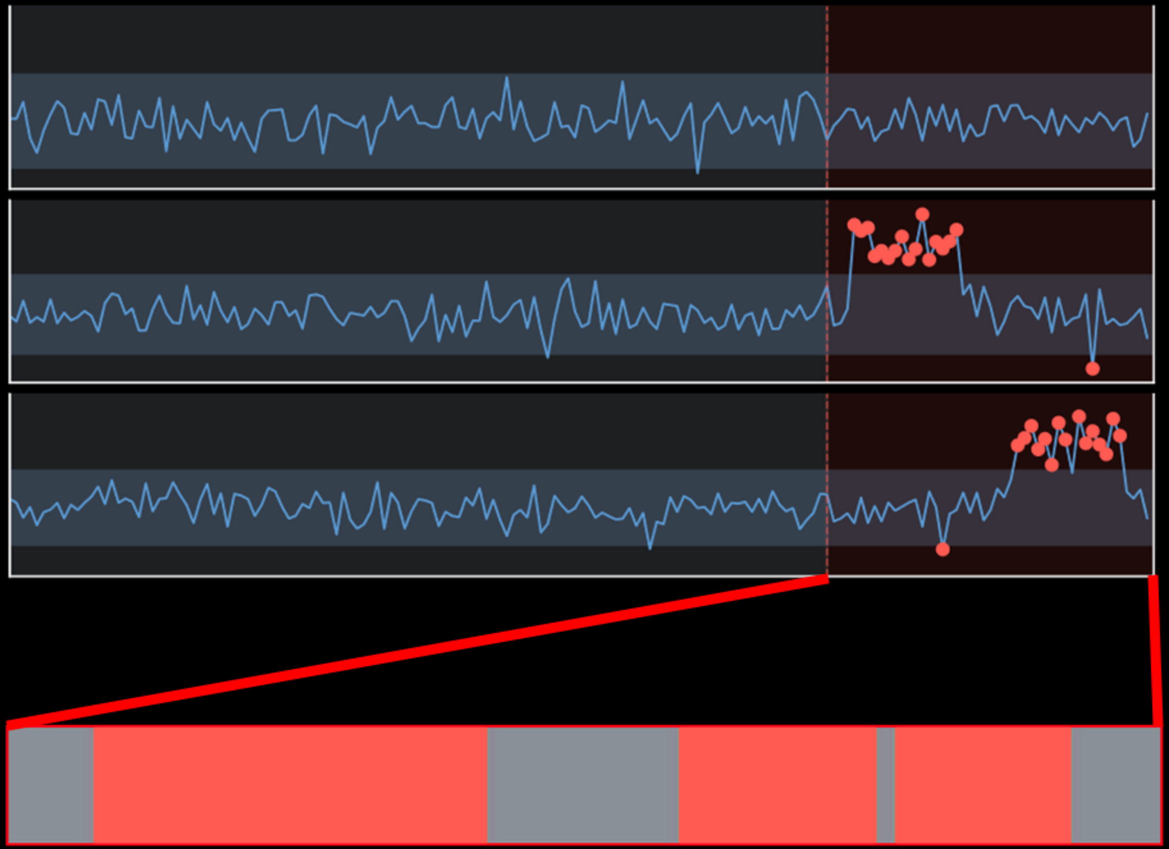
Univariate Test
& Univariate Ratio

Univariate Test

One channel carries the contextual anomaly

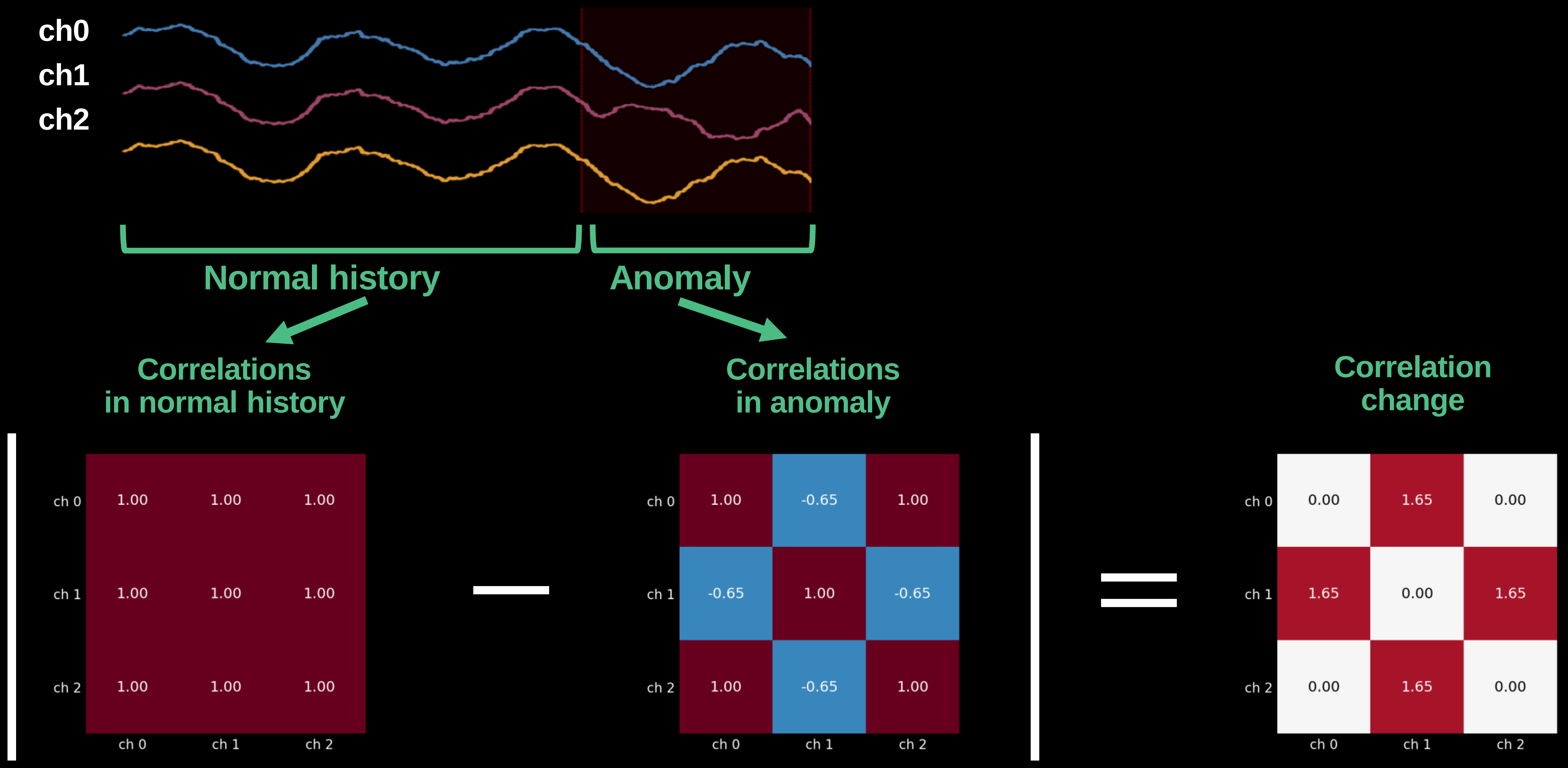
Univariate Ratio

Ratio of timesteps that trigger the univariate test in a contextual anomaly



31/48 anomalous timesteps = 0.65

Change in cross-channel structure



Classification results on these benchmarks

$\tau_z < 3$
$\tau_\rho > 0,1$

At reasonable thresholds,
no pure cross-channel anomalies

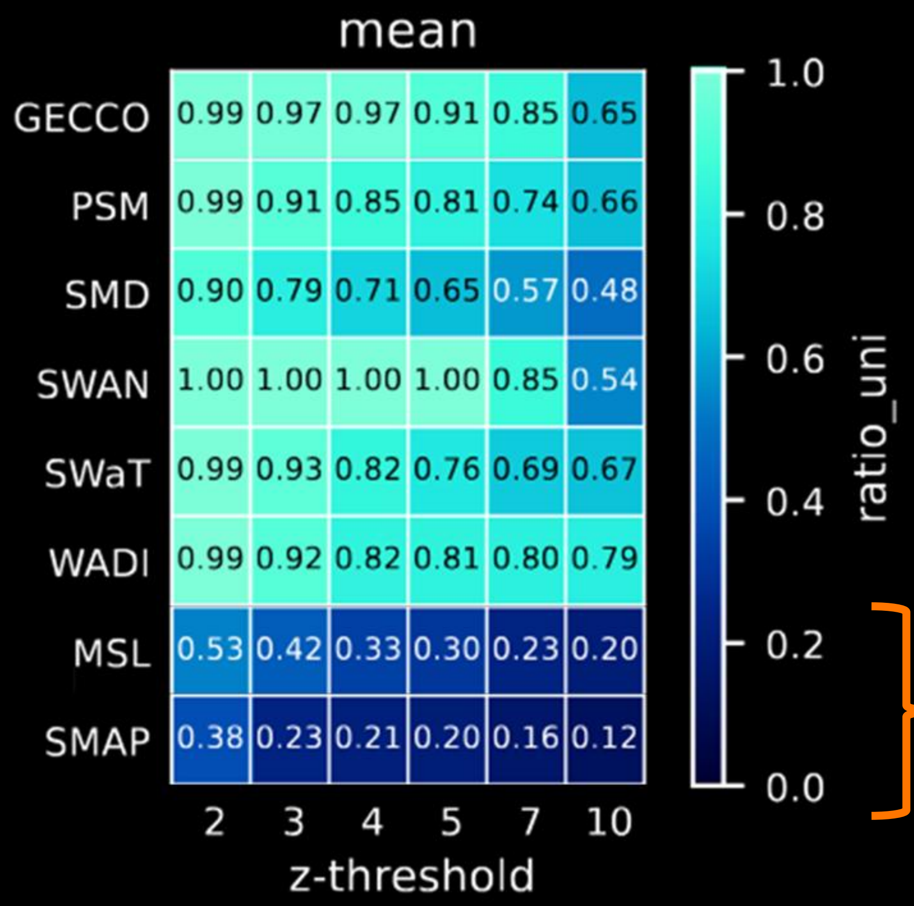
	CROSS-CH.			UNIVARIATE			BOTH			Short	
Dataset	P	S	D	P	S	D	P	S	D	UNI	UND
GECCO	0	0	0	0	0	0	22	22	22	0	0
MSL	0	0	0	4	3	31	30	31	3	0	0
SMAP	0	0	0	15	15	16	41	41	40	0	0
PSM	0	0	0	0	0	0	34	34	34	33	4
SMD	0	0	0	0	0	0	176	176	176	148	1
SWAN-SF	0	0	0	0	0	0	2	2	2	5079	1056
SWaT	0	0	0	0	0	0	35	35	35	0	0
WADI	0	0	0	0	0	0	14	14	14	0	0

Even at the most permissive thresholds,
only 9% of all segments across all benchmarks are purely cross-channel.

Univariate ratio: not a lone spike, univariate (almost) everywhere!

We evaluate how univariate the anomalies are

Even with absurd thresholds, most of them are still very univariate

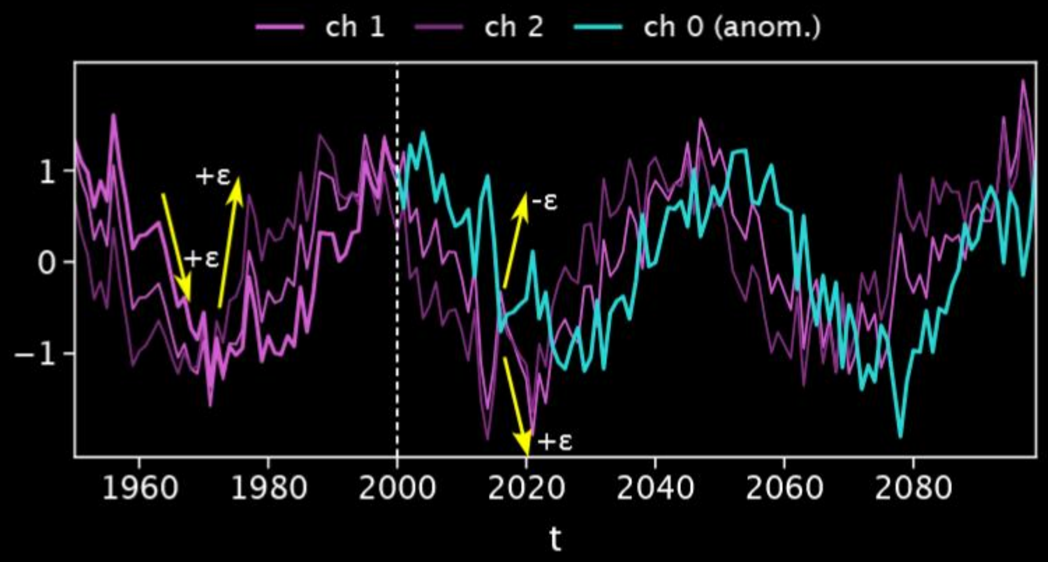


Avg. anomaly:
~80-100% univariate timesteps!

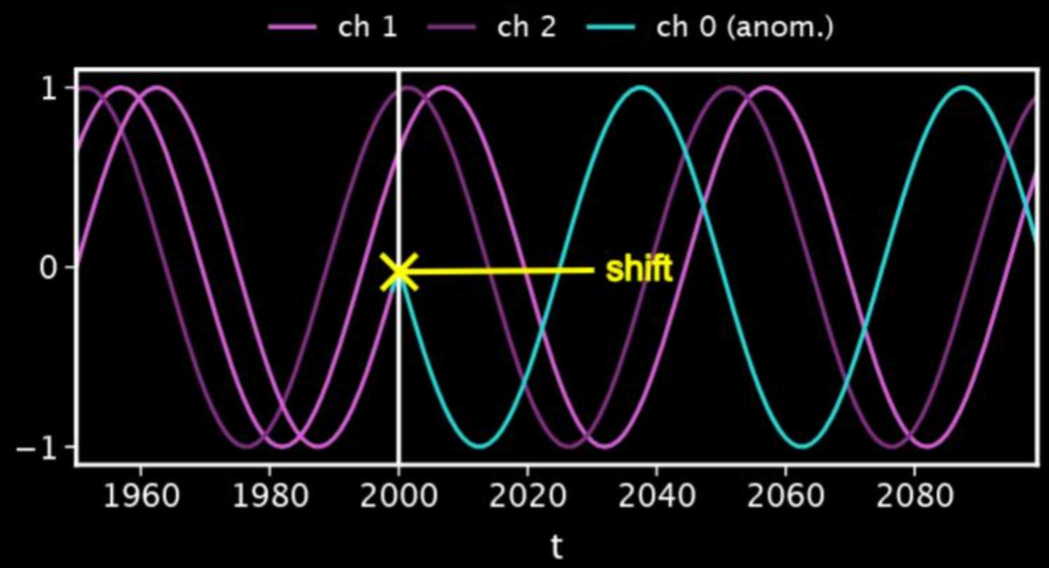
Binary channels → fallback to Univariate Matrix Profile
27/31 and 37/43 segments recovered (~87%)

But can our framework detect cross-channel anomalies when present?

Synthetic datasets with CROSS-CHANNEL-only anomalies



(a) NOISEFLIP.



(b) NPROLL.

noise removed for clearer view

Our framework correctly characterizes the synthetic anomalies!

Classification results out of 4000 segments

Dataset	CROSS-CH.			UNIVARIATE			BOTH		
	P	S	D	P	S	D	P	S	D
NOISEFLIP	3 967	3 967	3 967	0	0	0	33	33	33
NPROLL	3 970	3 970	3 970	0	0	0	30	30	30

99.2% success rate

Outliers due to Gaussian noise

No segment misclassified as UNIVARIATE

How does CD perform compared to CI on these synthetic data?

		NPROLL		NOISEFLIP	
Model	Var.	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Channel-dependent					
LinearAE	CD	0.97 ± 0.05	0.89 ± 0.16	0.84 ± 0.17	0.64 ± 0.35
AE	CD	0.97 ± 0.05	0.88 ± 0.16	0.81 ± 0.18	0.59 ± 0.33
CrossAD	CD	0.85 ± 0.09	0.61 ± 0.19	0.55 ± 0.04	0.21 ± 0.02
Channel-partiality					
CATCH	CP	0.77 ± 0.11	0.37 ± 0.20	0.66 ± 0.11	0.27 ± 0.12
Channel-independent					
LinearAE	CI	0.57 ± 0.06	0.17 ± 0.02	0.50 ± 0.01	0.15 ± 0.01
AE	CI	0.60 ± 0.06	0.18 ± 0.03	0.52 ± 0.02	0.16 ± 0.01
CrossAD	CI	0.53 ± 0.02	0.16 ± 0.01	0.47 ± 0.01	0.15 ± 0.01

Source: CrossAD: Time Series Anomaly Detection with Cross-scale Associations and Cross-window Modeling, Li et al. (NeurIPS 2025)
Source: CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching, Wu et al. (ICLR 2025)

How does CrossAD (CD) perform against (CI) on real benchmarks?

	VUS-ROC		VUS-PR	
Dataset	CI	CD	CI	CD
PSM	**0.7401**	0.7364	**0.5404**	0.5388
MSL	**0.8091**	**0.8091**	**0.3140**	**0.3140**
SMD	**0.8586**	0.8581	**0.2308**	0.2285
SMAP	**0.5784**	**0.5784**	**0.1444**	**0.1444**
GECCO	**0.9948**	0.7667	**0.6215**	0.0814
SWAN	**0.9554**	**0.9554**	**0.9278**	**0.9278**
SWaT	**0.6006**	0.5938	**0.3881**	0.3868

➔ **CrossAD (CD) detects cross-channel ruptures yet brings no gain over CrossAD (CI) on real benchmarks.**

Conclusion

Current benchmarks are not suited to test cross-channel capabilities

At reasonable thresholds, no strictly cross-channel anomalies in these benchmarks
They do appear at extreme ones but remain scarce ($< 9\%$ of the segments in all benchmarks)

When the anomaly is BOTH, it is still mainly univariate
The average anomaly deviates univariately on ~80-100% of its timesteps

➔ Our framework is reusable for any current and future MTSAD benchmark

➔ Our NPROLL and NOISEFLIP can be used to test cross-channel capabilities

Future Work

What to do next?

Development of more suited benchmarks to evaluate cross-channel modeling capabilities

Evaluate more datasets and focus on finding real cross-channel anomalies.

Advice for future benchmarks

No more channel anonymization!
If applicable, label per channel, not just per timestep

This could also pave the way for anomaly explanation…

Orange Research

Thank you!

Paper

Code

NPROLL Reconstruction Error

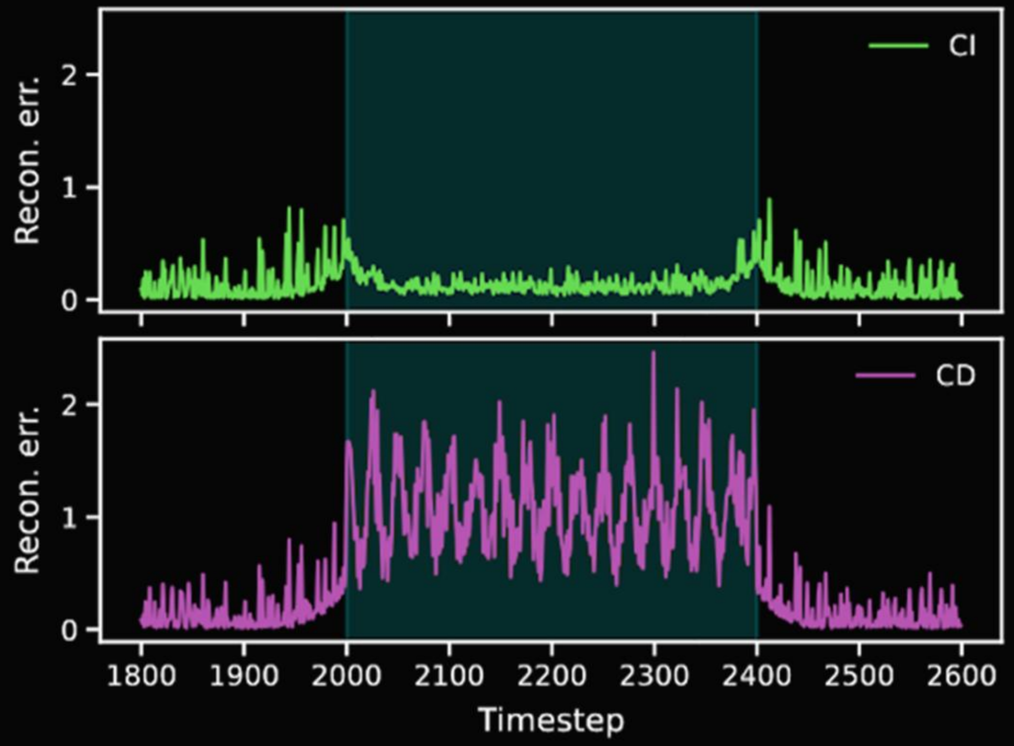
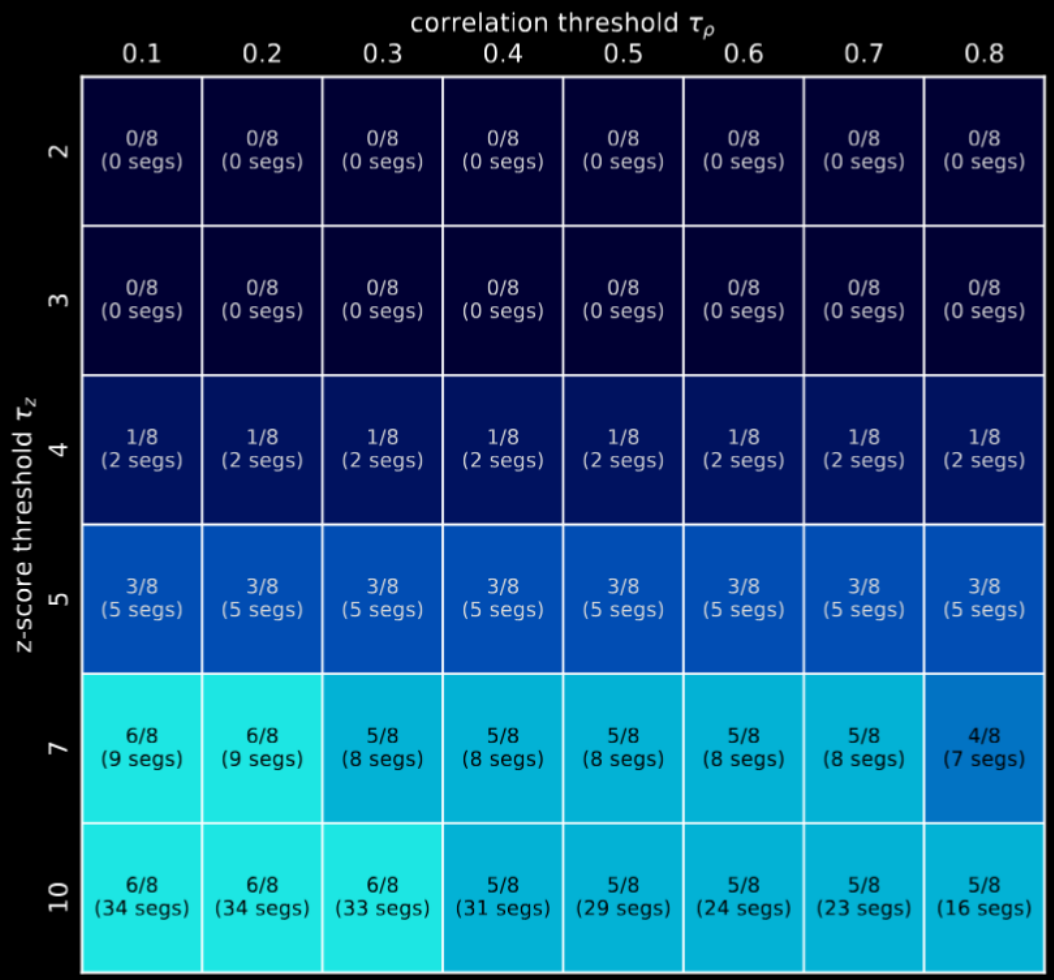


Figure 6: *Reconstruction losses from LinearAE (CI and CD) on an NPROLL anomaly segment. CI captures the boundaries but not the whole segment whereas CD fully captures it. Latent dimension is 16 and window size is 64.*

Anomaly characterization along two axes: threshold sensitivity



CROSS-CHANNEL only Anomaly (pearson)

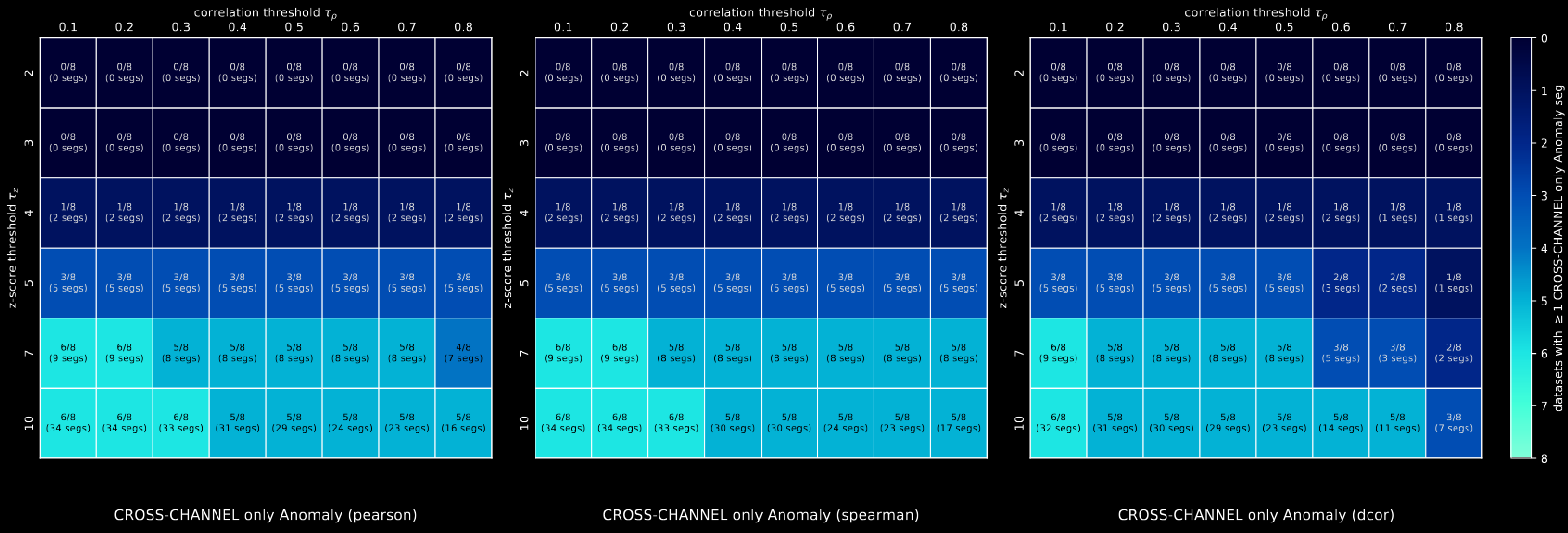
N/8 = how many benchmarks have a CROSS-CHANNEL anomaly

(… segs) = how many segments are CROSS-CHANNEL (sum on all benchmarks)

No CROSS-CHANNEL anomaly at reasonable thresholds

Even at the most permissive thresholds, only 9% of the segments for all benchmarks are CROSS-CHANNEL!

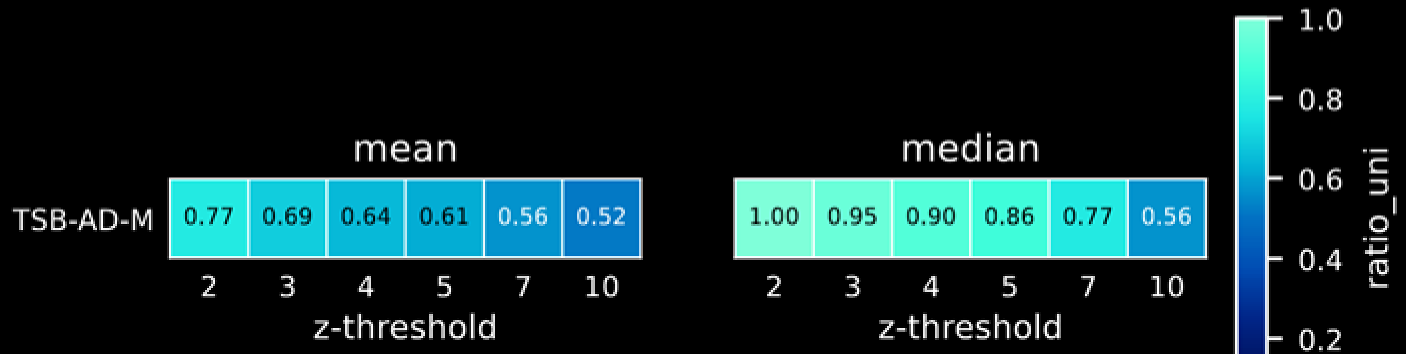
All correlation metrics



Number of CROSS-CHANNEL segments at different thresholds τ_z

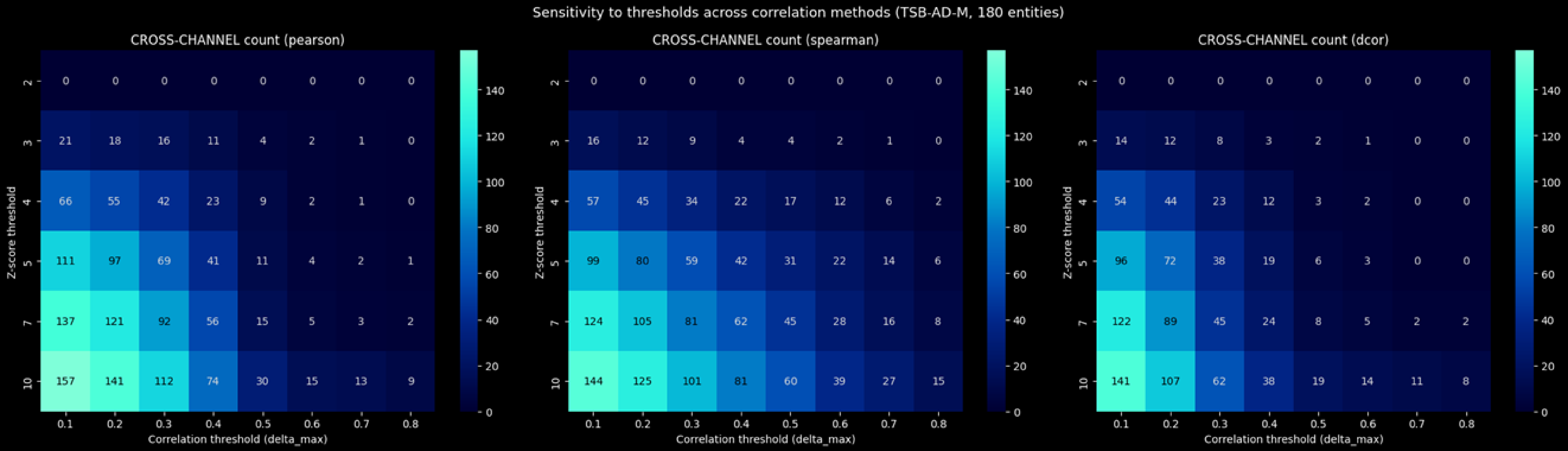
Run on TSB-AD-M

~9000 segments of length < 10 (excluded)



~3000 segments of length > 10:
- ~2500 univariate
- ~<u>500 both</u>

Only some CROSS-CHANNEL-only where $\tau_z = 3$



Number of CROSS-CHANNEL segments at different thresholds τ_z

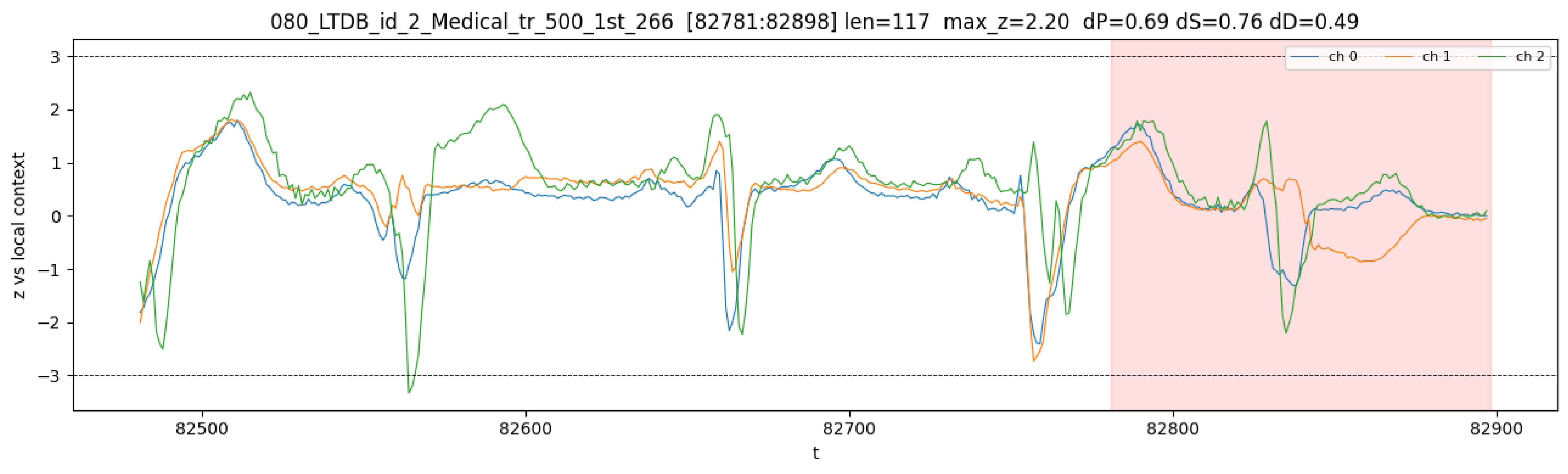
Run on TSB-AD-M: the only dataset that has cross-channel-only anomalies

Source: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, Goldberger et al. (2000)

VUS-PR (mean over all entities of LTDB)

LinearAE-CD : 0.3614
LinearAE-CI : 0.2093

Example



Formulas

$$z_c^{\mathcal{A}} = \max_{t\in\mathcal{A}} \frac{|X_{t,c} - \mu_c^{\mathcal{H}_{\mathcal{A}}}|}{\sigma_c^{\mathcal{H}_{\mathcal{A}}}}. \tag{1}$$

$$\Delta_\rho^{\max} = \max_{i\neq j} \left| R_{ij}^{\mathcal{A}} - R_{ij}^{\mathcal{H}_{\mathcal{A}}} \right|. \tag{3}$$

$$u^{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{t\in\mathcal{A}} \mathbf{1}\left\{ \max_{c\in\{1,\dots,C\}} \frac{|X_{t,c} - \mu_c^{\mathcal{H}_{\mathcal{A}}}|}{\sigma_c^{\mathcal{H}_{\mathcal{A}}}} > \tau_z \right\}. \tag{4}$$

$$X_{t,c} = \sin\left(\frac{2\pi t}{P} + \frac{2\pi c}{C}\right) + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \tag{5}$$

Default thresholds and Matrix Profile on MSL and SMAP

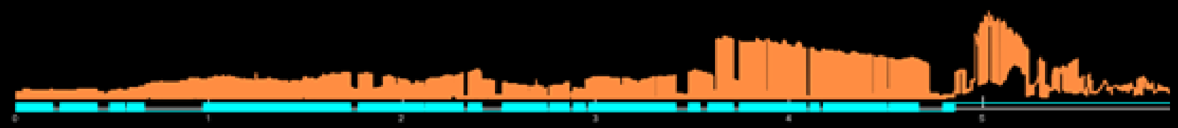
- $\tau_z = 3$
- $\tau_\rho = 0.1$
- $|\mathcal{A}| \geq 10$

MSL & SMAP: A Univariate Matrix Profile detected 64/74 segments *(after visual inspection, the remaining 10 were known labeling errors)*

Source: Current Time Series Anomaly Detection Benchmarks Are Flawed and Are Creating the Illusion of Progress, Wu & Keogh (TKDE 2021)

	UNIVARIATE			BOTH			Short	
Dataset	P	S	D	P	S	D	UNI	UND
GECCO	0	0	0	22	22	22	0	0
MSL	4	3	31	30	31	3	0	0
SMAP	15	15	16	41	41	40	0	0
PSM	0	0	0	34	34	34	33	4
SMD	0	0	0	176	176	176	148	1
SWAN-SF	0	0	0	2	2	2	5079	1056
SWaT	0	0	0	35	35	35	0	0
WADI	0	0	0	14	14	14	0	0

Most datasets are flawed: NIPS-TS-SWAN



This dataset has an extraordinarily high anomaly density. However, those apparently solid red bars are actually "spikes". Let's zoom in...

Source: Irrational Exuberance: Why we should not believe 95% of papers on Time Series Anomaly Detection, Eamonn Keogh (MILETS 2021)

Excerpts from the paper

3.1 Definitions

We denote by $\boldsymbol{X} \in \mathbb{R}^{T \times C}$ a multivariate time series of C channels and T timesteps indexed by $t \in \{1, \ldots, T\}$, with binary labels $\boldsymbol{y} \in \{0, 1\}^T$ where $y_t = 1$ marks an anomalous timestep.

An *anomalous segment* $\mathcal{A} = [t_s, t_e]$ is any maximal contiguous run of timesteps with $y_t = 1$ in $\boldsymbol{y}$. Its length is denoted $|\mathcal{A}| = t_e - t_s + 1$. Each segment is analyzed independently (referred to as *per-segment* throughout the paper).

For each anomalous segment $\mathcal{A}$, we define its *normal history* $\mathcal{H}_\mathcal{A}$ as the H closest normal timesteps preceding t_s. The history is constructed by walking backwards from t_s, skipping any anomalous timestep encountered along the way.

3.2 Univariate Test

For each anomalous segment $\mathcal{A}$, we estimate the per-channel normal mean $\mu_c^{\mathcal{H}_\mathcal{A}}$ and standard deviation $\sigma_c^{\mathcal{H}_\mathcal{A}}$ from $\mathcal{H}_\mathcal{A}$. We then compute, for every channel $c \in \{1, \ldots, C\}$, the maximum absolute z-score reached inside the segment:

$$z_c^\mathcal{A} = \max_{t \in \mathcal{A}} \frac{|X_{t,c} - \mu_c^{\mathcal{H}_\mathcal{A}}|}{\sigma_c^{\mathcal{H}_\mathcal{A}}}. \tag{1}$$

The segment is classified *univariate* if there is a channel $c \in \{1, \ldots, C\}$ such that $z_c^\mathcal{A} > \tau_z$. We use a default threshold $\tau_z = 3$. Indeed, Chebyshev's inequality guarantees that for any random variable X with finite mean μ and finite standard deviation σ:

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}, \quad \forall k > 0. \tag{2}$$

With $\tau_z = 3$, the probability that a single timestep deviates from its normal mean by more than $3\sigma_c^{\mathcal{H}_\mathcal{A}}$ under non-anomalous stationary conditions is bounded by $1/9 \approx 11\%$, regardless of the distribution (this is only an upper bound, not the actual probability). The threshold is therefore not chosen to fit any benchmark, and we report a sensitivity analysis over $\tau_z \in \{2, 3, 4, 5, 7, 10\}$ in Section 4.3.

While this condition is necessary for a segment to be classified as univariate, it is not sufficient on its own: the existence of a single timestep exceeding τ_z in one channel says nothing about how pronounced the univariate signal is across $\mathcal{A}$. We therefore complement this binary test with a per-segment intensity measure in Section 3.5.

We keep the z-score as the default test for its simplicity. When it is inapplicable, for instance on channels that are binary or nearly constant, we fall back to a Matrix Profile [27] applied on each channel independently, which captures unusual motifs rather than unusual values.

The z-score is deliberately a minimal univariate test, capturing only deviations in level relative to the recent history. Anomalies that the z-score misses but a richer univariate test would catch only strengthen our claim: they are still univariate.

Excerpts from the paper

3.3 Correlation Test

The cross-channel test compares the dependence structure between channels inside an anomalous segment with the one observed in its normal history. For each segment $\mathcal{A}$ and its history $\mathcal{H}_{\mathcal{A}}$, we compute two correlation matrices $\boldsymbol{R}^{\mathcal{A}}, \boldsymbol{R}^{\mathcal{H}_{\mathcal{A}}} \in \mathbb{R}^{C \times C}$ and quantify their discrepancy[3] by:

$$\Delta_{\rho}^{\max} = \max_{i \neq j} \left| R_{ij}^{\mathcal{A}} - R_{ij}^{\mathcal{H}_{\mathcal{A}}} \right|. \qquad (3)$$

The segment is classified *cross-channel* if $\Delta_{\rho}^{\max} > \tau_{\rho}$ for at least one of the three following correlation methods:

- the **Pearson** correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring linear dependence,
- the **Spearman** rank correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring monotone dependence,
- the **Distance correlation** [28], in its unbiased squared form [14] $\rho_{ij} = \mathrm{dCor}_U^2 \in [-1, 1]$ (negative values are bias-correction noise, clipped to zero [29]), which captures any form of statistical dependence and is harder to trigger than Pearson/Spearman at high τ_{ρ}.

The threshold $\tau_{\rho} \in [0, 2]$ controls the sensitivity of the cross-channel test. We use a permissive default threshold $\tau_{\rho} = 0.1$ and report a sensitivity analysis in Section 4.3. We use three measures rather than dCor alone because dCor is computed at lag zero only, while Pearson and Spearman allow an efficient lagged extension that captures lead-lag dependencies.

Extension to lag Channel pairs in real-world systems often exhibit time-shifted relationships [30]. To capture such dependencies, we extend Pearson and Spearman to lagged correlations: $\hat{\rho}_{ij}(\ell)$ measures, on the training set, the correlation between channel i at time t and channel j at time $t + \ell$, for $\ell \in \{-L_{\max}, \dots, L_{\max}\}$.

Let $\mathcal{L} = \{-L_{\max}, \dots, L_{\max}\} \setminus \{0\}$ denote the set of non-zero lags. A pair (i, j) is recorded as *lagged-dominant* if $\exists \ell_{ij}^{\star} \in \mathcal{L}$ such that $|\hat{\rho}_{ij}(\ell_{ij}^{\star})| > |\hat{\rho}_{ij}(0)|$. When computing $\boldsymbol{R}^{\mathcal{A}}$ and $\boldsymbol{R}^{\mathcal{H}_{\mathcal{A}}}$ in Equation (3), the entry (i, j) for any lagged-dominant pair uses the lagged correlation at $\ell_{ij}^{\star}$ rather than at lag zero.

Distance correlation is intentionally not extended to lags. Computing dCor_U^2 at every lag for every channel pair was prohibitive even on the smallest benchmark in both channel count and series length.

Excerpts from the paper

Excerpt, a reproduced page of a paper:

3.3 Correlation Test

The cross-channel test compares the dependence structure between channels inside an anomalous segment with the one observed in its normal history. For each segment $\mathcal{A}$ and its history $\mathcal{H}_{\mathcal{A}}$, we compute two correlation matrices $\boldsymbol{R}^{\mathcal{A}}, \boldsymbol{R}^{\mathcal{H}_{\mathcal{A}}} \in \mathbb{R}^{C \times C}$ and quantify their discrepancy[3] by:

$$\Delta_\rho^{\max} = \max_{i \neq j} \left| R_{ij}^{\mathcal{A}} - R_{ij}^{\mathcal{H}_{\mathcal{A}}} \right|. \tag{3}$$

The segment is classified *cross-channel* if $\Delta_\rho^{\max} > \tau_\rho$ for at least one of the three following correlation methods:

- the **Pearson** correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring linear dependence,
- the **Spearman** rank correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring monotone dependence,
- the **Distance correlation** [28], in its unbiased squared form [14] $\rho_{ij} = \mathrm{dCor}_U^2 \in [-1, 1]$ (negative values are bias-correction noise, clipped to zero [29]), which captures any form of statistical dependence and is harder to trigger than Pearson/Spearman at high τ_ρ.

The threshold $\tau_\rho \in [0, 2]$ controls the sensitivity of the cross-channel test. We use a permissive default threshold $\tau_\rho = 0.1$ and report a sensitivity analysis in Section 4.3. We use three measures rather than dCor alone because dCor is computed at lag zero only, while Pearson and Spearman allow an efficient lagged extension that captures lead-lag dependencies.

Extension to lag Channel pairs in real-world systems often exhibit time-shifted relationships [30]. To capture such dependencies, we extend Pearson and Spearman to lagged correlations: $\hat{\rho}_{ij}(\ell)$ measures, on the training set, the correlation between channel i at time t and channel j at time $t + \ell$, for $\ell \in \{-L_{\max}, \ldots, L_{\max}\}$.

Let $\mathcal{L} = \{-L_{\max}, \ldots, L_{\max}\} \setminus \{0\}$ denote the set of non-zero lags. A pair (i, j) is recorded as *lagged-dominant* if $\exists \ell_{ij}^* \in \mathcal{L}$ such that $|\hat{\rho}_{ij}(\ell_{ij}^*)| > |\hat{\rho}_{ij}(0)|$. When computing $\boldsymbol{R}^{\mathcal{A}}$ and $\boldsymbol{R}^{\mathcal{H}_{\mathcal{A}}}$ in Equation (3), the entry (i, j) for any lagged-dominant pair uses the lagged correlation at ℓ_{ij}^* rather than at lag zero.

Distance correlation is intentionally not extended to lags. Computing dCor_U^2 at every lag for every channel pair was prohibitive even on the smallest benchmark in both channel count and series length.

Univariate ratio at tau_z = 3

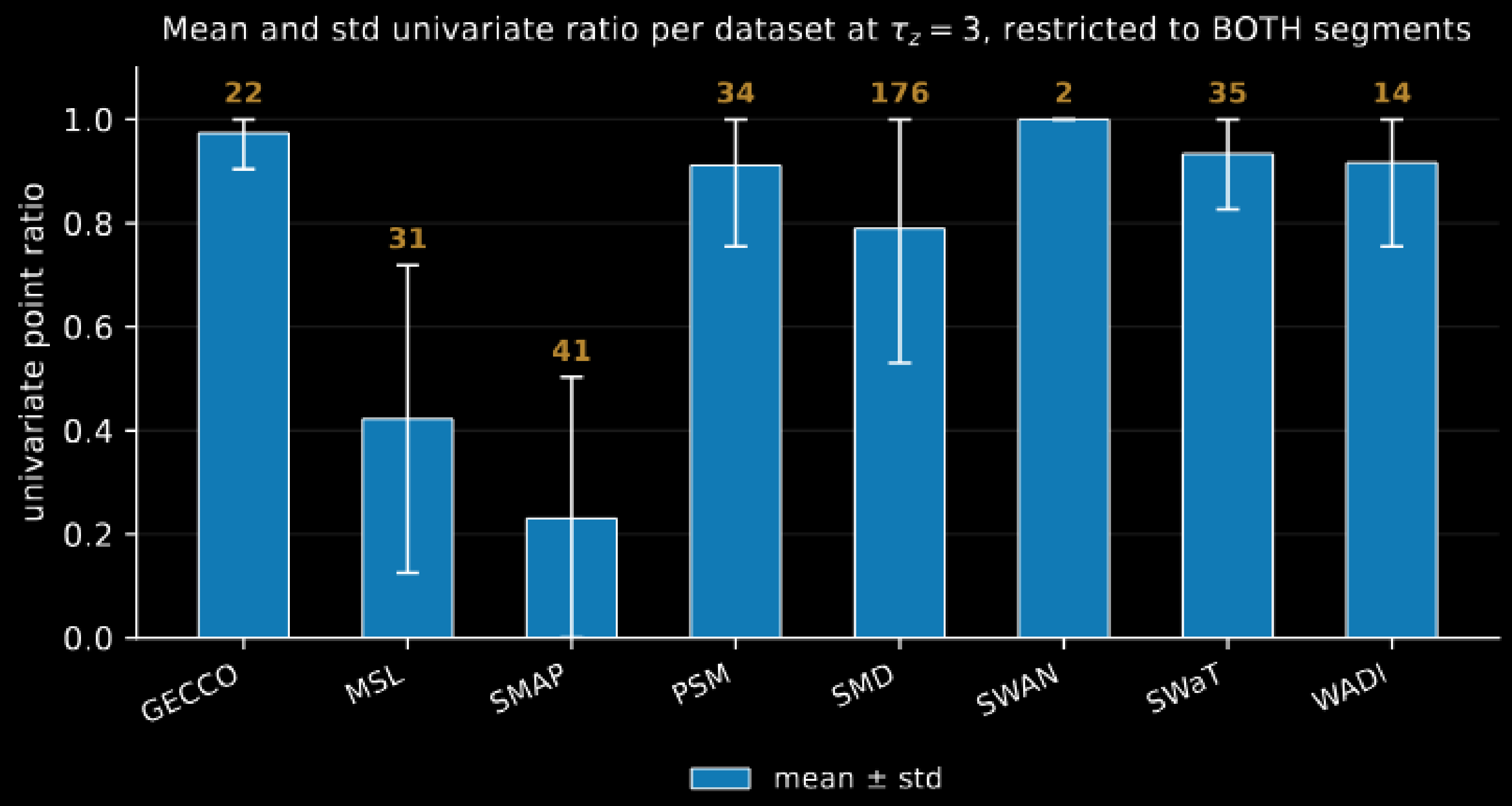


***Figure 3:** Mean univariate ratio (with standard deviation) per dataset at $\tau_z = 3$, restricted to BOTH segments. The number above each bar is the segment count.*

Univariate ratio (mean and median)