



Towards explainable anomaly detection for satellite telemetry

Schefels Clemens¹ · Balan Arvind¹ · Ben Salem Bilel² · Gerhardus Andreas² · Helmsauer Kathrin¹ · Lambert Baptiste² · Niebling Julia² · Rewicki Ferdinand² · Rings Thorsten² · Schlag Leonard¹

Received: 15 April 2026 / Revised: 23 July 2026 / Accepted: 20 August 2026
© The Author(s) 2026

Abstract

The increasing complexity of modern satellites and the growing amount of telemetry data available pose significant challenges for a safe and economic operation of satellites. To support the satellite engineers, traditional machine learning methods, including deep learning-based approaches, have shown promising results but lack intuitive explainability, hindering their adoption in operational settings. This paper presents a novel approach to anomaly detection and causal inference in satellite telemetry data, leveraging an ensemble of classical statistical models and deep learning architectures, combined with causal discovery techniques. We employ the Peter and Clark momentary conditional independence algorithm for identifying causal relationships with temporal dependencies and use its results as part of an in-depth root cause analysis enhancing anomaly detection. Our approach identifies potential anomalies and provides indications which satellite components cause the detected anomalies in order to facilitate interpretation by satellite operators. By integrating causal inference methods into anomaly detection pipelines, we aim to enhance explainability and facilitate decision-making in complex systems. This paper contributes to the growing body of work on anomaly detection and causal inference, highlighting the potential of combining machine learning and causation for improved operational performance.

Keywords Anomaly detection · Causal discovery · Machine learning · Satellite operations · Correlation analysis · Signal processing

1 Introduction

When operating satellites in space, telemetry data serve as a critical resource, enabling engineers to monitor the satellite's health and system status. As spacecrafts become increasingly complex and the number of satellites in orbit is increasing, the amount of telemetry data available for analysis is growing as well. A single modern satellite with a scientific payload defines and transmits around 80,000 different telemetry parameters to Earth. Since the manual inspection and analysis of every parameter is not feasible anymore, many space operations centers have established research groups focused on investigating and developing anomaly detection methods, including novel approaches, to address this challenge.

The German Space Operations Center (GSOC)¹ at the German Aerospace Center (DLR)² has developed ATHMoS, the Automated Telemetry Health Monitoring System [1], which is currently used by satellite engineers to aid in detecting novel behavior [2–4]. It is based on classical machine learning methods, using statistical features of the satellite telemetry [5], and, by design, returns highly explainable results to the operators [6, 7]. The current state of research in this domain is heavily focusing on deep-learning-based approaches and has shown promising results on benchmark datasets. However, a significant limitation of these approaches is their lack of intuitive explainability which makes it difficult to include them into the daily work routine of satellite operators. Without understanding the cause of a detected novelty and why it was even detected, operators are limited in their ability to take appropriate action and lose trust in detection results.

✉ Schefels Clemens
clemens.schefels@dlr.de

¹ German Space Operations Center (GSOC), German Aerospace Center (DLR), Münchener Straße 20, 82234 Weßling, Bavaria, Germany

² Institute of Data Science, German Aerospace Center (DLR), Mälzerstraße 5, 07745 Jena, Thuringia, Germany

¹ <https://www.dlr.de/rb/>.

² <https://www.dlr.de/>.

To address this problem, a collaborative initiative between GSOC and the DLR Institute of Data Science (DW)³ as part of the DLR project CausalAnomalies⁴ seeks to uncover novel insights in satellite telemetry by combining anomaly detection methods with causal inference techniques, see, e.g., [8–11]. Unlike ATHMoS, which focuses on individual telemetry parameters, this case study uses high-dimensional data combining satellite telemetry from multiple parameters. In this investigation, we employ both deep learning and classical machine learning methods, such as isolation forest, one-class support vector machine, cluster-based local outlier factor (CBLOF), graph-augmented normalizing flows (GANF), and maximally divergent intervals (MDI) algorithm.

Given the recent progress in anomaly detection and attribution by merging methods from machine learning and causal inference [12–17], we here explore the potential of such approaches for satellite systems. The additional application of a causal inference method adds more value to these “black-box” or “semi-transparent” methods like Bayesian or graph neural networks [18, 19], variational auto-encoder [20], and evidential convolutional neural networks [21] by providing causal explanations for the detected anomalies. Without this extra step, many of these detection methods would lack viability to be used operationally. By finding causal relationships and utilizing them in a root cause analysis, we aim to make the detection results easily interpretable by the satellite engineers and operators.

This paper builds upon our prior analysis [22] of satellite telemetry data, in which we examined ground station contact events to identify and explain anomalous behavior. The findings and operational insights from this analysis enabled us to systematically refined our anomaly detection and causal inference techniques. These advancements were applied to a new use case involving an Earth observation satellite operated by GSOC.

This paper is structured as follows: in Sect. 2, we report on the diverse methodologies for detecting and explaining anomalies in multivariate time series. Specifically, our ensemble approach employs classical statistical models and deep learning-based architectures for unsupervised anomaly detection. Furthermore, as part of our causal analysis, we infer the causal structure between the system variables and use this information for root cause analysis. Section 3 is dedicated to our use case, the detection and attribution of anomalies in satellite telemetry data. We investigate data collected during satellite image acquisition (SIA) events, and aim to understand the underlying causes of said anomalies. Here, various telemetry data related to SIA events are

analyzed, including status, temperature, power, and attitude control data. Concluding with Sect. 4, we provide a final summary of our results as well as an overview of planned future activities.

2 Methods

In this section, we introduce and describe the methods we use for our data analysis. This subsumes methods from the fields of anomaly detection and causal inference.

2.1 Anomaly detection on unlabelled data

Selecting and calibrating an anomaly detection model is challenging when working with unlabelled data, especially in complex systems where anomalies are difficult to characterize or when working with large datasets for which accurate labeling by experts is infeasible, such as in our use case in Sect. 3. To overcome this challenge, we propose a three-step iterative framework that leverages an ensemble of anomaly detection models, aiming to systematically identify the most relevant anomaly candidates for subsequent investigation. This framework is outlined in Fig. 1.

2.1.1 Step 1: preprocessing

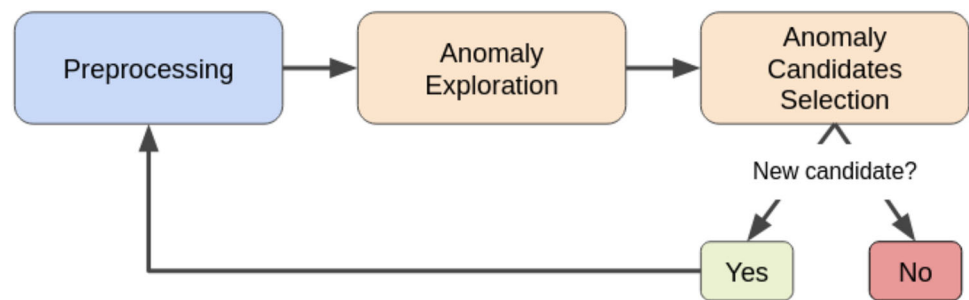
We analyze multivariate time series data given by the dataset $\mathcal{X} = \{X^{(1)}, \dots, X^{(N)}\}$ consisting of N observations $X^{(k)} \in \mathbb{R}^{D \times T_k}$ where D denotes the number of variables and T_k the number of potentially irregularly spaced time steps in observation k . The number of variables per multivariate time series is constant across all observations in $\mathcal{X}$. Where necessary, $X_i^{j(k)}$ denotes the value of the i th variable at the j th time step in observation k . For brevity, we will omit superscripts and subscripts where they are not needed.

To make the irregular time series data suitable for anomaly detection models, we resample each observation onto a regular grid of 200 points. This step ensures a consistent number of time steps per observation and guarantees that each variable has an assigned value at every grid point. For each variable and grid segment (i.e., between two consecutive grid points), the mean, standard deviation, and the number of data points in the segment is computed.

To reduce the dimension of the multivariate time series, we select a single representative variable among each set of highly correlated variables. We denote two variables as highly correlated if their absolute Pearson correlation coefficient $|\rho|$ exceeds 0.99. Rather than training a dedicated anomaly detection model, potential anomalies associated with these variables are identified directly by comparing pairs of highly correlated variables X_m and X_n , $m, n \in$

³ <https://www.dlr.de/dw/>.

⁴ <https://www.dlr.de/de/forschung-und-transfer/projekte-und-missionen/causal-anomalies/>.

Fig. 1 Process for identifying anomaly candidates

$\{1, \dots, D\}$. Consistent with the procedure described in Sect. 2.1.2, observations with an anomaly score exceeding three standard deviations from the mean are flagged as anomalous. The correlation-based anomaly score is defined as

$$\|X_m - \text{sign}(\text{corr}(X_m, X_n))X_n\|_2,$$

where $\text{corr}(\cdot, \cdot)$ is Pearson's correlation coefficient and $\|\cdot\|_2$ is the Euclidean norm. Potential anomalies identified in this manner are subsequently incorporated into the anomaly candidates list introduced in Sect. 2.1.3.

Following best practices, the dataset $\mathcal{X}$ is split into training and test subsets $\mathcal{X}_{\text{train}}, \mathcal{X}_{\text{test}} \subseteq \mathcal{X}$. Because our aim is to identify all potentially anomalous observations, $\mathcal{X}_{\text{train}}$ is drawn as a random subset of unique observations $X^{(k)}$ with its size being at most 80% of the size of $\mathcal{X}$. Suspected anomalies identified in Sect. 2.1.3 are excluded from any training subsets to improve model quality. If more than 20% of observations are flagged as anomalous, the iterative process is terminated since anomalies are expected to be rare. Although this did not occur in our analysis, a high anomaly rate would indicate model miscalibration and require restarting the procedure with a more restrictive threshold than the current choice of three standard deviations above the mean anomaly score.

2.1.2 Step 2: anomaly exploration

In the next step, we detect anomalies in the preprocessed data. As the absence of labels makes it difficult to assess the quality of a single anomaly detection method directly, we rely on an ensemble of five anomaly detection methods of various types to detect a wide range of different anomalies as suggested in [23]:

- *Isolation forest* [24]: a tree-based model that isolates anomalies by randomly selecting features and splitting values.
- *One-class support vector machine (SVM)* [25]: a SVM that learns a decision boundary around normal data points.

- *Cluster-based local outlier factor (CBLOF)* [26]: a context-aware outlier detection method, capturing both isolated anomalies and small groups of atypical points relative to larger cluster structures, by leveraging k -means clustering and local outlier factors.
- *Graph-augmented normalizing flow (GANF)* [27]: a density-based method that leverages normalizing flows for density estimation and integrates a Bayesian Network to model causal relationships among multivariate time series.
- *Maximally divergent intervals (MDI) algorithm* [28]: a density-estimation-based algorithm for unsupervised detection of sequential anomalies in spatio-temporal data.

For all models, we fix the input window size to 200 time steps, according to the grid definition in Sect. 2.1.1. Model-specific thresholds are set to three standard deviations above the mean anomaly score computed on the training data.

2.1.3 Step 3: anomaly candidates selection

While building an ensemble model from five distinct anomaly detection methods helps mitigating the limitations of individual models, it does not guarantee all anomalies to be detected. To increase coverage, 100 models are trained for each type using different hyperparameter configurations, all of which are then included in the ensemble. An observation $X^{(k)}$ is then flagged as a potential anomaly based on a majority vote among the total of 500 model instances. This multi-instance ensemble strategy mitigates the problem of missing targets for hyperparameter tuning and allows us to detect a wide range of anomalies, even when certain model instances fail to recognize specific anomaly types, based on the findings of [23].

The model hyperparameters could be selected randomly; however, we choose to generate them via Bayesian optimization that maximizes the model confidence defined by the separation between the mean anomaly score of predicted normal data and that of predicted anomalies. Although the objective is not strict model optimization, this procedure biases the hyperparameter distribution toward configurations

that produce more robust models with clearer distinction between normal and anomalous data.

At the end of this step, the new list of potential anomalies is compared with anomalies identified in previous iterations of this process, as well as expert-labeled anomalies (if applied). If no new anomalies are detected, the process terminates. Otherwise, the newly identified anomalies are added to the anomaly candidates list, and the procedure is repeated from Step 1, see Sect. 2.1.1.

2.2 Causal analysis

As second part of our data analysis methodology, we employ methods from causal inference. This field of research develops and applies methods for reasoning about causal relationships data, see, e.g., [8–11] for elaborate accounts.

2.2.1 Causal discovery

Within causal inference, *causal discovery* or alternatively *causal structure learning* refers to inferring qualitative cause-and-effect relationships from data. That is, given data of a collection of variables $X_i, i \in \{1, \dots, d\}$, the goal is to learn for all $1 \leq a \neq b \leq d$ whether or not X_a has a causal influence on X_b . This binary information can be conveniently represented by a graph, often referred to as *causal graph*, in which directed edges signify direct causal influences; for example, an edge $X_a \rightarrow X_b$ signifies that X_a has a direct causal influence on X_b . For an overview of foundations of as well as algorithms for causal discovery, see, e.g., [9, 11, 29, 30].

When applying causal discovery or interpreting its results, it is important to keep the following well-known fact in mind: statistical association does not necessarily imply causation. Attempting to infer causal relationships from observational data *alone* is, thus, an ill-posed problem. However, the problem can become well-posed if appropriate assumptions on the data-generating process are adopted. These enabling assumptions differ between the several approaches and various algorithms for causal discovery.

One type of causal discovery algorithms follow the so-called *constraint-based* approach. This approach utilizes (in-)dependencies between variables in the data. For example, if the causal graph is $X_1 \leftarrow X_2 \rightarrow X_3$, then one expects that X_1 and X_3 are marginally dependent (because their common cause X_2 introduces a dependence) and independent conditional on X_2 (because the dependence is exclusively due to X_2). Thus, if the data do not show these two (in-)dependencies, then we would want to reject the above causal graph. Then, the idea is to “reverse-engineer” the causal graph (or, more precisely, a set of possible causal graphs) from patterns of (in-)dependencies in the data. One algorithm that operationalises this strategy is the Peter and Clarke

(PC) algorithm [31]. Since, in practice, the independencies are not known a priori, they must be tested using statistical independence tests applied to the data.

Due to the principled distinction between associational and causal relationships, this reverse engineering can only provably work if appropriate assumptions are imposed. For the PC algorithm these are: First, the assumption of acyclic causal relationships. Second, the so-called *causal sufficiency* assumption [9], which says that there are no unobserved variables that causally influence two observed variables. Third, the so-called *causal Markov* assumption [9], which says that every variable is independent of its non-descendants given its parents as specified by the causal graph. Fourth and last, the *causal faithfulness* assumption [9], which says that every independence in the data is implied by those independencies that are due to the causal Markov condition.

In the work presented here, we employ the multi-data-version [32] of the Peter–Clark Momentary Conditional Independence Plus algorithm (PCMCI⁺) [33], which is available in the Python package *Tigramite*.⁵ The PCMCI⁺ algorithm [33] is an adaption of the PC algorithm to multivariate discrete-time time series, and its multi-data-version allows to pool data from multiple multivariate time series (here observations), assuming that these are multiple instances of the same process. Unlike the PCMCI algorithm [34], on which PCMCI⁺ is based, PCMCI⁺ (and its multi-data-version) also allows for contemporaneous causal relationships, i.e., causal influences between different variables at the same time step—as long as these are acyclic. This ability to handle contemporaneous relationships is vital in this study as the variables potentially interact on time scales shorter than the employed sampling rate.

We combine this algorithm with the *RegressionCI* independence test as implemented in the *Tigramite* package,⁶ which is similar to the test described in [35]. This test can handle time series containing both continuous and discrete variables as is the case here. Neglecting some details, to test whether X_a and X_c are independent given X_b , the test works by first regressing X_a on X_b only as well as jointly on X_b and X_c , and subsequently test whether the two fits differ significantly. Here, the form of the regression in the first step and of the test in the second step depend on which of the involved variables are continuous and which are discrete. We reject statistical independence of two variables at a significance level of $\alpha = 0.05$ (unless stated otherwise) and allow for delayed causal relationship up to $\tau_{\max} = 2$ time steps. As the samples for independence testing are drawn in a sliding-window fashion from the time series, stationarity of the time

⁵ Available at: <https://github.com/jakobrunge/tigramite/tree/master/tigramite>; version 5.2.7.0.

⁶ See https://github.com/jakobrunge/tigramite/blob/master/tigramite/independence_tests/regressionCI.py; version 5.2.

series is implicitly assumed. This implicit assumption might be a source of errors as stationarity is not guaranteed in our data.

Here, we aim to infer a “common causal graph” G_{com} that represents the satellite system across all observations. This aim is the reason for using the multi-data-version of PCMC⁺, as we can pass multiple observations to the algorithm and have it infer one causal graph across. In order for this inferred graph to be as reliable as possible, we employ the following strategy: First, we inform the algorithm about several causal relationships that are known to exist thanks to expert knowledge; forcing the corresponding edges to be part of the causal graph simplifies the overall inference task.

Second, we test the consistency of the inferred causal graphs by pooling different numbers v of observations and randomly selecting $w = 20$ different subsets. For each set of inferred graphs $\{G_u\}_v$, $v \in \{10, 20, 50\}$, $u \in \{1, \dots, w\}$, we compare how strongly the graphs differ from each other using the *structural hamming distance (SHD)* [36]. Given two graphs with the same vertex set, this distance counts the minimal number of edge additions, deletions, and reversals (that is, replacing $X_a \rightarrow X_b$ by $X_a \leftarrow X_b$; this counts as a single change) with which the two graphs can be made equal.

Third, we refine the set of graphs $\{G_u\}_v$ with the smallest average SHD by pruning the graphs of edges that are inconsistent with expert knowledge. We additionally remove edges that do not exist in all graphs of the set, effectively building the intersection of the graphs, starting with edges with the weakest indication of existence (smallest test statistics of the conditional independence test). However, if removing an edge would disconnect the graph, then we refrain from pruning it. This yields the common causal graph G_{com} .

2.2.2 Root cause analysis

Root cause analysis (see, e.g., [13, 37]) is based on the idea that, in a complex system, anomalous events can spread from one component to another along the pathways of causal influences. Consequently, it might be that the “root cause” of an anomaly can be found in a different component than the one in which the anomaly is first detected. To advance our understanding of encountered anomalies, we can leverage the knowledge gained by causal discovery to identify root causes by tracing influences “upwards” through the causal structure of the system. We follow and adapt the work of Janzing et al. [38] as well as Budhathoki et al. [12], who discuss pointwise anomalies in the form of singular outliers among exogenous influences on variables. We extend their approach to anomalous time series (here observations flagged by our anomaly detection approach; “anomalous observation” for future reference) instead of pointwise outliers.

To this end, we require the following three assumptions: first, the causal relationship between the variables of the

system are known (e.g., in the form of a causal graph); second, the variables follow a structural causal model (see, e.g., [8, 11, 39]): each variable X_i can be expressed as a function $f_i(\text{PA}(X_i), U_i)$ of its causal parents $\text{PA}(X_i)$ —as specified by the causal graph—and a noise term U_i that models exogenous influences, such that the noise terms $U_1, \dots, U_d$ are jointly independent of each other. Here, the joint independence of noise terms corresponds to the assumption of causal sufficiency [9] in causal discovery. Note, however, that the functions f_i need not be known; third, anomalies do not affect the causal structure of the system (i.e., the set of causal parents remain unaltered).

With these assumptions, we employ so-called *conditional z-scores* to investigate which variables are root causes of anomalies and which are only inheriting them from their causal parents. Usually, the z-score is a measure of deviation of a data point of a variable X_i from the mean of its distribution in units of the distribution’s standard deviation:

$$z_i = \frac{X_i - \mathbb{E}[X_i]}{\sqrt{\text{Var}[X_i]}},$$

where $\mathbb{E}[\cdot]$ is the expectation value (here estimated by the arithmetic mean) and $\text{Var}[\cdot]$ is the variance. This process is also known as standardizing and is often employed to compare variables with widely separated ranges of values. By conditioning the expectation value and variance on the causal parents of X_i , *conditional z-scores* z_i effectively remove the influence of $\text{PA}(X_i)$ on X_i :

$$\tilde{z}_i = \frac{X_i - \mathbb{E}[X_i | \text{PA}(X_i)]}{\sqrt{\text{Var}[X_i | \text{PA}(X_i)]}}. \quad (1)$$

To estimate the conditional expectation value $\mathbb{E}[\cdot | \cdot]$ and conditional variance $\text{Var}[\cdot | \cdot]$, we here employ a k -nearest neighbor regressor (number of neighbors = 5 and uniform weights; see, e.g., [40]) as we cannot expect linear relationships between the different variables. Importantly, our third assumption allows us to regress on all samples independent of anomalies. We also explicitly incorporate causal links at delays τ —as specified by the time-lagged causal relationships—by including relevant time series time-shifted by τ in the regressions. We end up with time series of conditional z-scores $\{\tilde{z}_i\}_t$ since the conditional expectation values and variances reflect the time dependence of the various conditions. These time series, then, can be used as input of our anomaly detection method and the following downstream analysis.

2.2.3 Generalized ordinal pattern-based outlier detection

Instead of using information theory-based outlier scores and counterfactuals for singular outliers [12, 38], we employ an ordinal pattern analysis [41, 42] to better identify variables

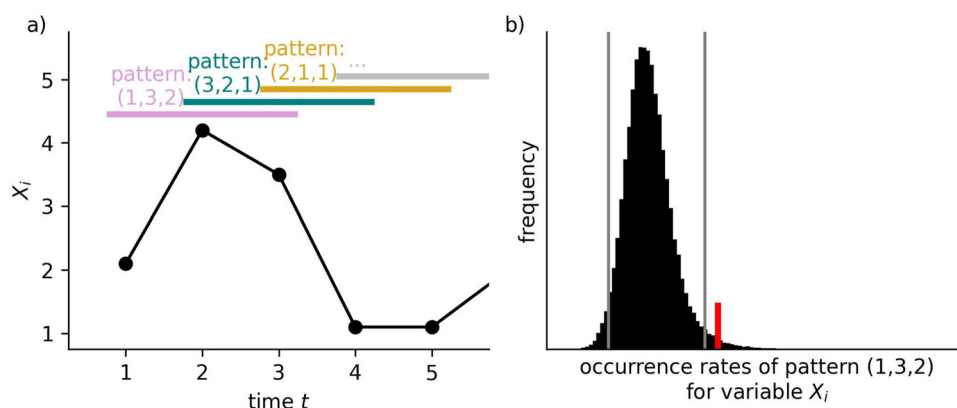


Fig. 2 Toy example of a generalized ordinal pattern transformation and root cause candidate identification. **(a)** The time series of variable X_i is transformed into patterns of embedding dimension $m = 3$. Single patterns are indicated by colored vertical bars at the top (magenta, green, yellow, silver), while the correspondingly colored text annotates the rank representation of the pattern (three dots indicate continuation beyond the border of the figure). As an example, the first pattern (magenta) is described by the sequence of low value (rank 1), high value (rank 3), and intermediate value (rank 2). **(b)** Toy example distribution

as root cause candidates that show contextual deviations. Specifically, we transform both the time series $\{X_i\}_t$ and $\{\tilde{z}_i\}_t$ separately into sequences of so-called generalized ordinal patterns [43]. Generalized ordinal patterns are completely defined by position and rank of data points with respect to each other and can handle ties of subsequent values, see Fig. 2a for an example. The length of the pattern (number of symbols in a pattern) is called embedding dimension m , which we set to 3. A lower embedding dimension (e.g., $m = 2$) would result in simple monotonicity checks, while a higher dimension results in too sparsely sampled patterns given the number of data points per observation. The chosen embedding dimension produces 13 individual patterns, which reflects the ordered Bell number [43].

Technically, we derive the generalized ordinal patterns by densely ranking the values of $\{X_i\}_t$, resp. $\{\tilde{z}_i\}_t$, within moving windows of length m . We select the windows so that they overlap, and with successive windows shifted by one data point relative to their predecessors.

Following this transformation, we count the occurrence rate of each pattern in each variable and in each normal observation. This results in 13 distributions, one per individual pattern, over the observations per variable X_i . For anomalous observations, we then check if patterns from the different variables are outside the 1- to 99-percentile range of the empirical distributions of the normal observations, see Fig. 2b. If at least one pattern of a given variable is outside this range, we consider the variable a root cause candidate.

Finally, we check against randomness. Since we are now testing for outliers of occurrence rates of patterns multiple

of occurrences of pattern (1, 3, 2) in a variable X_i from normal observations and a single anomalous observation. Gray vertical lines indicate the 1st and the 99th percentile of the distribution, respectively. The red vertical line indicates the occurrence rate of the pattern from an anomalous observation. Since the rate is outside the prescribed range of the distribution of normal observations, we consider this an indication of X_i being a root cause as the patterns are based on conditional z -scores and thusly are corrected for the effect of causal parents

times ($\epsilon = 13$; the number of patterns), we have to make sure that we do not draw these outliers by chance. To that end, we employ the hypergeometric test (see, e.g., [44]) for each anomalous observation and each variable. Assuming randomness of occurrence rates, the probability of identifying one outlier outside of the 1- to 99-percentile range is $p^{\text{outlier}} = 0.02$. Assuming independence between the 13 patterns, the probability of encountering at least γ outliers of patterns in a single variable of an anomalous observation by chance is

$$p^{\text{chance}} = \sum_{\eta=\gamma}^{\epsilon} \binom{\epsilon}{\eta} (p^{\text{outlier}})^{\eta} (1 - p^{\text{outlier}})^{\epsilon-\eta}. \quad (2)$$

If this probability is below a significance level of $\alpha = 0.05$, we consider the root cause candidate significant. However, since we test multiple anomalous observations this way, we also have to correct our significance level with the Bonferroni correction by the number of variables times the number of tested anomalous observations.

2.2.4 Sensitivity analysis

Since our downstream methods depend on correctly identified causal structures, we evaluate their performance under structural errors. Inspired by network surrogates [45], we perturb a subset of edges in G_{com} to generate 20 erroneous graphs, denoted $G_{\text{com}}^{(s)}$ for $s \in [1, \dots, 20]$.

For each perturbed graph, we randomly remove 5% of the edges and reconnect them between randomly selected vari-

ables. Throughout this process, we preserve the total number of directed edges, ensure all perturbed graphs remain directed acyclic, and exclude edges derived from expert knowledge from randomization. Additionally, if a removed edge represents a temporal delay, the replacement edge maintains the same lag.

Using these perturbed graphs, we compute modified conditional z -scores $\{\tilde{z}_i\}_t^{(s)}$. These scores serve as inputs for the anomaly detection method in Sect. 2.1 and the generalized ordinal pattern-based outlier detection in Sect. 2.2.3.

To quantify method sensitivity, we first compare the anomaly detection results using $\{\tilde{z}_i\}_t^{(s)}$ against those obtained from the original conditional z -scores $\{\tilde{z}_i\}_t$ (computed from G_{com}). Second, we verify whether root cause analysis yields consistent results when applied to time series flagged as anomalous using $\{\tilde{z}_i\}_t^{(s)}$. For both comparisons, we use the F1 score, treating the anomaly labels resp. root causes from the original z -scores $\{\tilde{z}_i\}_t$ as the ground truth.

3 Use cases

In our prior work [22], we conducted a detailed analysis of satellite telemetry data to assess ground station contact events, with a focus on identifying and interpreting anomalous interactions using causal inference. Our investigation encompassed a wide array of telemetry parameters critical to maintaining communication integrity and diagnosing operational irregularities. Key parameters included signal strength metrics for low-rate and high-rate receivers, alongside down-link status indicators. We also examined transmission relay states for both high-rate and low-rate transmitters to ensure proper communication pathways during ground contacts. The satellite's power distribution system, including BUS electricity supply and its allocation between subsystems and payload functions, was thoroughly analyzed.

Building on the findings and operational insights from this study, we refined our telemetry analysis framework and causal inference techniques. These improvements were then applied to a new use case:

Earth observation satellites are instrumental in monitoring our planet and its resources, offering critical insights for environmental management, climate change research, disaster response, and sustainable development. GSOC operates several earth observation satellites, one of which serves as the primary resource for this research.

However, satellites are intricate systems comprising multiple components that work together to achieve their objectives. These include the camera system, reaction wheels, temperature sensors, power systems, and attitude control systems—each playing a vital role in the satellite's performance. Any issues with one component can have a ripple effect throughout the entire system.

Table 1 List of satellite telemetry variables. Labels, signal types, and descriptions of parameters of the satellite system employed as variables in this study

Label	Type	Description
P01	Status	Camera
P02	Current	Power processing unit
P03	Current	Payload
P04	Temperature	Power processing board
P05, P06	Temperature	Payload platform 1, 2
P07	Status	Commanded attitude mode
P08	Status	Current attitude mode
P09	Status	Current attitude sub mode
P10	Degree	BIAS roll
P11	Degree	BIAS pitch
P12	Status	Indicator S/C in eclipse
P13–P16	Rotation rate	Reaction wheel 1–4
P17	Power	Main power BUS
P18, P19	Charge	Battery 1, 2
P20	Current	Battery charge
P21	Current	Solar array load
P22	Current	S/C load
P23	Temperature	Battery stack
P24	Temperature	SBC worker
P25	Temperature	SBC monitor
P26	Temperature	Mid solar panel

Fortunately, telemetry data from these components hold immense value for optimizing satellite operations. By analyzing these data, we can identify anomalies and predict potential failures, enhancing the overall efficiency and reliability of satellite operations.

The earth observation satellite telemetry data used in this study are collected during satellite image acquisition (SIA) events. As the satellite captures images at regular intervals, it provides a continuous source of data that offers insights into the satellite's performance. These data include information on the satellite's status, power consumption, temperature, and attitude. In detail, we analyze the following parameter data (see also Table 1 and Fig. 3):

- *Status* (P01, P12): The parameters display the status of the camera on board and indicate when the satellite is in eclipse.
- *Reaction wheels* (P13–P16): These parameters record the rotation rate of the reaction wheels.
- *Temperature* (P04–P06, P23–P26): These parameters provide temperature readings from different parts of the satellite.
- *Power* (P02, P03, P17–P22): These parameters monitor the satellite's power system, including batteries and solar panels.

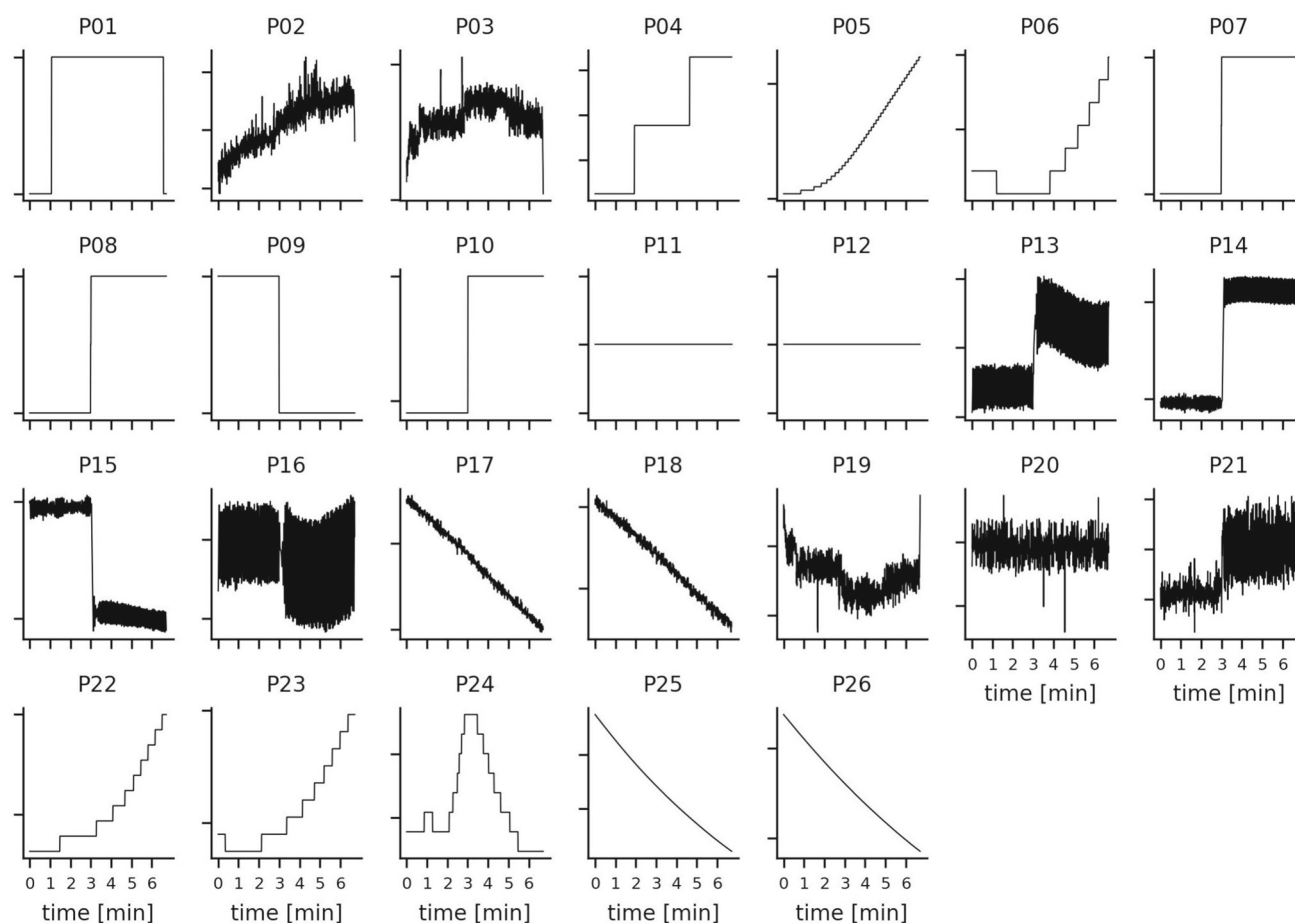


Fig. 3 Exemplary time series of satellite telemetry data recorded during a SIA. During a SIA, a multitude of sensors and computer components of a satellite system advise on the condition of the spacecraft. Of these readings, the depicted 26 variables have been selected by domain experts

on the basis of their high information content in relation to our study. They consist of both categorical (e.g., P01) and continuous (e.g., P02) variables. For the sake of comprehensibility, all time series are depicted in appropriate units

- *Attitude control* (P07, P08, P09, P10, P11): These parameters provide information on the satellite's attitude, such as its roll, pitch, and yaw angles.

The primary objective of this study is to detect anomalies during SIA events and elucidate their underlying causes. To achieve this, we employ machine learning-based anomaly detection techniques to identify deviations from expected behavior that may indicate potential system failures or performance degradation. Furthermore, we utilize causal inference methods to relate detected anomalies and telemetry parameters, thereby providing a deeper understanding of potential failure modes and operational inefficiencies. By detecting and interpreting these anomalies, we aim to facilitate proactive measures to mitigate potential issues before they impact satellite operations. Ultimately, our findings will provide valuable insights that can inform future satellite design and operational strategies, leading to improved robustness and resilience.

3.1 Data processing

To ensure accurate anomaly detection and causal analysis, we perform a series of preprocessing steps on the raw telemetry data. First, we load all available telemetry records and use the camera status parameter to identify the start and end times of each SIA event. The camera status serves as a reliable indicator of when the camera is active, providing natural boundaries for segmenting relevant telemetry data. We then extract the corresponding telemetry parameters around each identified SIA event, including an additional buffer period of 3 min before and after the event. This buffer allows us to capture transitional behaviors and potential precursors to anomalies, providing a more comprehensive understanding of system dynamics during these events. The result is a structured multivariate time series dataset that represents telemetry data from 690 observed SIA events, forming a robust foundation for our subsequent analysis.

Table 2 List of the SIAs with the highest anomaly scores. In the columns time series and cond. z-score, we marked the 10 SIAs with the highest anomaly score, where for column time series, we used the raw time series as input for the anomaly detection methods while for column cond. z-score, we put in the conditional z-scores derived from the time series. The SIA labeled abnormal by a majority of models is marked with *. Column expert labels shows the expert feedback

SIA #	time series	cond. z-score	expert labels
28		*	x
43		x	
45		x	x
47	x		
81	x		
90	x	x	x
91	x	x	x
94		x	x
95	*	x	x
145	x		
161	x		x
167	x		x
395		x	x
396		x	
570		x	
614	x		
623	x		

This final step is followed by further processing for anomaly detection (Sect. 3.2), causal analysis (Sect. 3.3), and root cause analysis (Sect. 3.4).

3.2 Anomaly detection

Following the process described in Sect. 2.1, we trained and evaluated 100 distinct hyperparameter configurations for each of the five selected anomaly detection models at each cycle, using either the time series data or the conditional z-scores as input. After three cycles, no additional anomaly candidates were identified, indicating convergence of the iterative process. In total, 17 anomaly candidates were detected out of the 690 SIA we analyzed as reported in Table 2. We illustrate two examples of variables exhibiting atypical behavior during these anomalous SIAs in Fig. 4.

In addition to these candidates, we identified three SIAs (#38, #85, and #86) exhibiting pronounced de-correlations within groups of otherwise highly correlated variables.

While anomaly detection models are suited to find outliers, they are not providing any insights on the nature of the anomaly and potential threats to the system. Therefore, each of the anomaly candidates found by our anomaly detection process were also evaluated by domain experts to assess whether or not the outliers are of concern for the system. If the

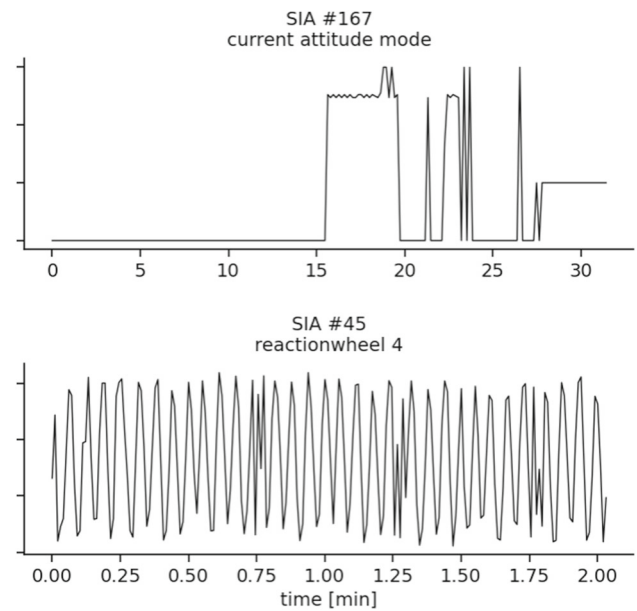


Fig. 4 Two examples of variables exhibiting atypical behavior during anomalous SIAs. Anomalous SIAs have been identified with our machine learning approach (see Sect. 2.1), and depicted variables were selected with a counterfactual explanations method sensible to atypical behavior. In comparison to normal behavior (cf. Fig. 3), we found sharp single-observation spikes for P08 during SIA #167. During the anomalous SIA #45, the rotation rate P16 is constantly oscillating at high frequency and average value. For the sake of comprehensibility, all time series are depicted in appropriate units

expert labels are considered as ground truth, the conditional z-score-based approach is the superior preprocessing strategy with an F1 score of 0.48, while the time series-based anomaly detection approach achieves an F1 score of 0.34, as shown in Table 3. Generally, results show that additional information is required to distinguish benign outliers from anomalies of concern, a task that can be supported by methods from causal inference and root cause analysis.

3.3 Inferring the causal structure

In order to further improve our understanding of the nature of the anomalies detected in the previous section and the system in general, we next inferred causal relationships between the telemetry data variables and checked for the stability of the inferred relationships. With the help of expert knowledge, we enforced 27 causal relationships when employing multi-data PCMC⁺ and derived three sets of causal graphs $\{G_j\}_m$ (see Sect. 2.2).

We encountered the largest difference between causal graphs for datasets of 10 SIAs, which was expected as statistical accuracy tends to increase with the number of data points. On average, the structural hamming distance (SHD) between each pair in the $\{G_j\}_{10}$ set was 13.7. By incorporating more SIAs, the distance reduced to 9.4 for $\{G_j\}_{20}$ and to

Table 3 Performance comparison of the proposed ensemble approach to its best anomaly detection child models on the conditional z -score and time series datasets

	Model	F1	PR_AUC	Precision	Recall
Cond. Z-score	LOF	0.154	0.200	0.250	0.111
	IFOREST	0.200	0.319	0.113	0.889
	CBLOF	0.432	0.251	0.286	0.889
	GANF	0.462	0.299	0.353	0.667
	OCSVM	0.026	0.013	0.013	1.000
	MDI	0.069	0.133	0.038	0.333
	Ensemble	0.480	0.520	0.700	0.370
Time series	LOF	0.105	0.127	0.100	0.111
	IFOREST	0.188	0.372	0.105	0.889
	CBLOF	0.286	0.135	0.175	0.778
	GANF	0.320	0.367	0.250	0.444
	OCSVM	0.026	0.013	0.013	1.000
	MDI	0.148	0.089	0.089	0.444
	Ensemble	0.340	0.460	0.500	0.260

7.3 for $\{G_j\}_{50}$. To further increase confidence in the stability of inferred relationships, we also varied the significance level of the statistical independence test ($\alpha \in \{0.1, 0.05, 0.01\}$) with only minor influence on the resulting graphs (data not shown), and we continued with $\alpha = 0.05$.

Given that the most consistent results were achieved with 50 SIAs, we further refined the graphs of $\{G_j\}_{50}$. We pruned on average 11.6 superfluous edges from each $G_j \in \{G_j\}_{50}$ with additional expert knowledge, excluding unrealistic causal relationships. This way, we attained a single “common causal graph” G_{com} , which is depicted in Fig. 5.

Interestingly, we saw only weak indication for the existence of the enforced causal relationships (low values of test statistics of conditional independence tests; cf. edge colors in Fig. 5). Since many of these edges relate categorical variables reflecting setpoints and other controls (e.g., P08: current attitude mode) with continuous variables [e.g., P13: rotation rates (reaction wheel 1)], this portends to an issue with the RegressionCI independence test when handling categorical variables with little temporal variability.

3.4 Root cause analysis

In the next step, we employ the root cause analysis (RCA) approach described Sect. 2.2.2 to identify possible candidates of system components ultimately responsible for the anomalous behavior of the flagged anomalous SIAs. We leverage our generalized ordinal pattern analysis and outlier statistics with both the original data and the conditional z -scores as differences between the resulting sets of root cause candidates are expected to provide insight into the propagation of the identified anomalies. Figure 6 shows which variables are considered root cause candidates in which anomalous SIA.

For the original time series data, we find that all but five (P07, P09, P10, P11, P22) of the considered variables are

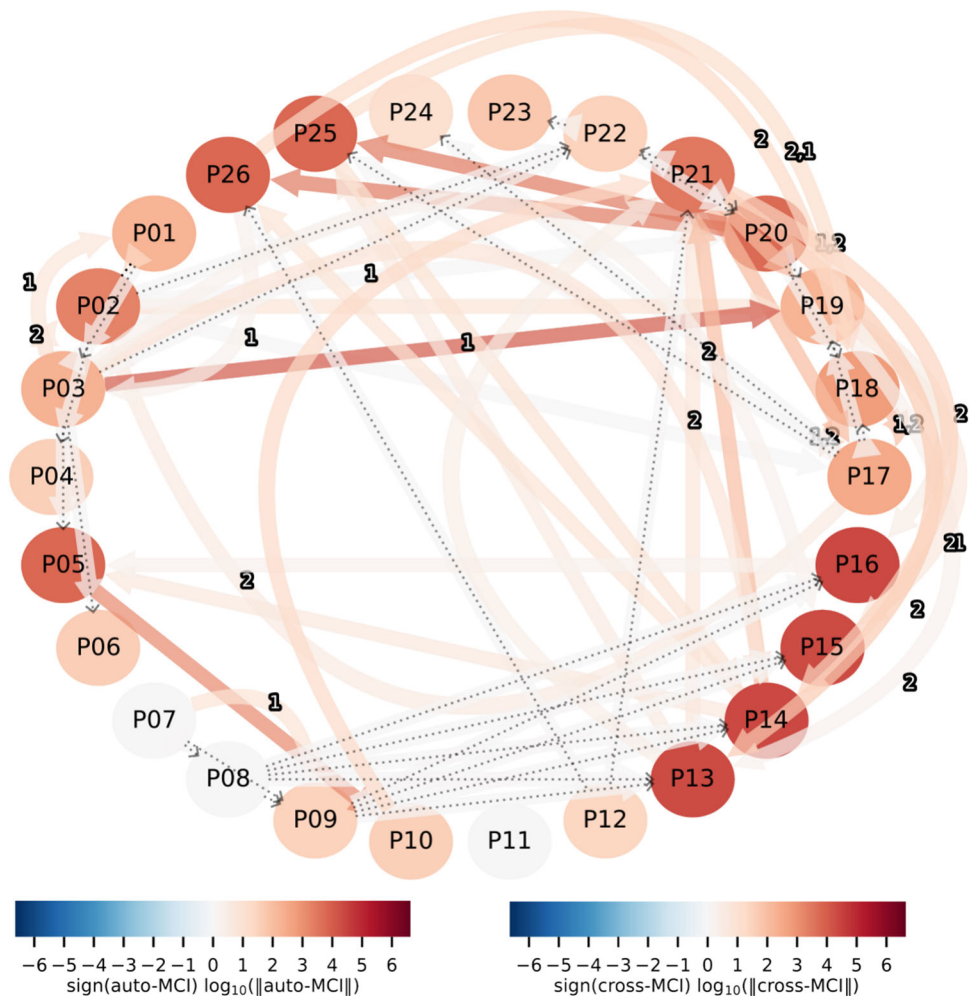
identified as root cause candidates in at least one of the anomalous SIAs. There are 5.0 ± 3.7 (mean and standard deviation; range 0–11) candidates per anomalous SIA. These candidate variables were also all significant in at least one of the anomalous SIAs, and there are, on average, 3.5 ± 3.1 (range 0–9) significant candidates per anomalous SIA. Notably, we cannot identify any root cause candidates with our approach in three anomalous SIAs—namely #396, #614, and #623—when investigating the original data.

For conditional z -scores, we generally find fewer root cause candidates and these candidates are also less often significant, but we find at least one candidate in each anomalous SIA. Again, only the same five variables (P07, P09, P10, P11, P22) are never candidates and we have 4.1 ± 3.7 (range 1–11) candidates per anomalous SIA. Aside from P02, these candidates were also all significant in at least one of the anomalous SIAs (mean and standard deviation of significant candidates per anomalous SIA: 2.8 ± 3.1 ; range 0–9).

Focusing on the significant root cause candidates, we can have four different results: the first is straightforward: a variable is not a significant candidate in either the original time series data or the conditional z -scores (e.g., P01 in SIA 28; cf. Fig. 6). Here, we can reasonably assume that the dynamics of the variable is normal; the second case is the opposite. The variable is considered a significant candidate in both the original data and in the conditional z -scores (e.g., P12 in SIA 43). In this case, we can infer that the variable is a true root cause: it behaves anomalous and the behavior can not be traced back to a causal parent.

The two remaining case are more involved. In the third case, we find a variable to be a significant candidate in the original data but not in the conditional z -scores (e.g., P02 in SIA 81). In this case, we can assume that the anomalous behavior of the variable is inherited from a causal parent, who are either anomalous themselves or appear normal (com-

Fig. 5 Common causal graph G_{com} of the telemetry data acquired with the multidata PCMC⁺ algorithm from 20 sets of 50 SIAs, including expert knowledge and pruning. Vertices represent the investigated telemetry data variables. Straight edges represent contemporaneous causal relationships between variables and point from causal parent to child variables. Curved edges depict relationships at a non-zero delay indicated by the accompanying black-rimmed numbers (in units of the sampling rate of the data). Edge color shows the value of the test statistic of the employed conditional independence test (color scale at the lower right). Node color shows the same but for auto-independence, i.e., between variables and their own past (color scale at the lower left). Gray dotted cores of edges indicate causal relationships that were identified by expert knowledge and fixed during causal discovery



binning information from Figs. 5 and 6). The latter can be the result of a compound effect (multiple parents show only weak, undetectable anomalous behavior, which add up in the child variable) or circumstantial (the behavior is normal for each single parent, but should not happen in concert with the other parents' behavior). Finally, the fourth case happens when a variable is identified as a significant candidate in the conditional z -scores but not in the original data (e.g., P16 or P18 in SIA 396). In this case, we have to assume that the effect of the parents on the variable obfuscates anomalous behavior in the original data, meaning that the variable is indeed a root cause. We conjecture that case three and four are part of the reason why we find different anomalies with our machine-learning-based anomaly detection approach when we investigate either the original data or the conditional z -scores.

3.4.1 Detailed analysis of a SIA event

To illustrate the practical utility of our approach, we present a detailed analysis of a single SIA event. We focus on SIA #95, which was flagged as anomalous with the highest confidence

score by our model, i.e., the top-10 anomaly score on conditional z -scores and majority vote on raw time series. It is confirmed by our satellite engineers (Table 2) too. Our root cause analysis identified 13 candidates, 11 significant, using raw time series data and 9 candidates, all significant, using conditional z -scores (cf. Fig. 6).

The satellite engineer confirmed the anomalous nature of this SIA based on the following observations (Fig. 7):

1. The image acquisition duration was unexpectedly short (visible in P01; data not shown), suggesting an early termination likely caused by a software issue.
2. The thermal sensors monitoring of the payload platform (P04–P06: camera and support structure) exhibited a flat linear increase rather than the expected increasing steepness (cf. Fig. 3).
3. The reaction wheels (P13–P16) oscillated at low amplitudes, indicating an attempt to stabilize the satellite's orientation rather than execute a deliberate adjustment. This was corroborated by the constant status of the attitude mode variables (P07–P09; data not shown). While

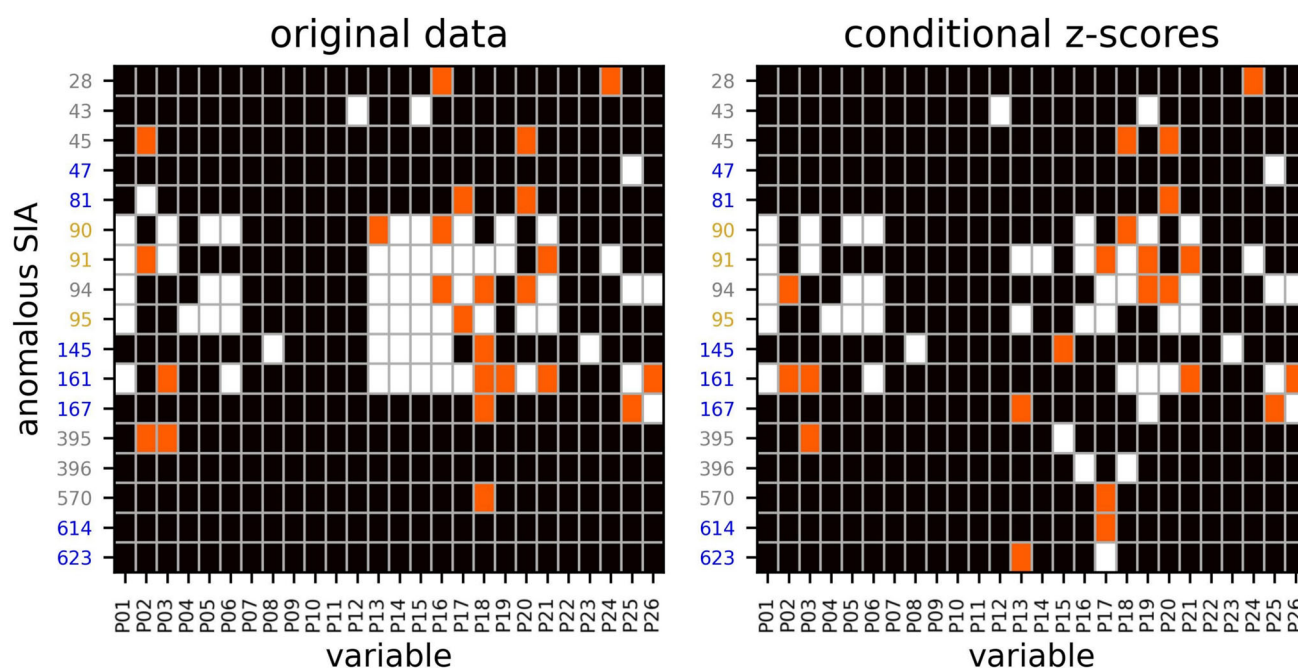


Fig. 6 Root cause candidates for each anomalous SIA based on a generalized ordinal pattern analysis of the original data (left) and conditional z -scores (right). Pixel color indicates which variables are considered root cause candidates (orange) and which candidates can also not be explained by chance and, hence, are considered significant (white; hypergeometric test, $p < 0.05$, Bonferroni corrected). Black pixels indicate behavior of original data or conditional z -scores that do not

deviate from their behavior during normal SIAs according to ordinal pattern analysis (see Sect. 2.2). The color of the SIA labels indicate whether or not conditional z -scores were used as input for our machine learning-based anomaly detection scheme: anomalies were found with the original data (blue), with the conditional z -scores (gray), or with both data sets (gold)

constant values in individual attitude mode variables are technically within normal operating bounds (as the satellite does not always require orientation changes), the satellite engineer identified simultaneous constancy across all attitude mode variables as anomalous. Also note that the conditional z -scores correctly remove the anomalous behavior from two of the four reaction wheel variables (a candidate variable of the type of *case three*).

4. The remaining flagged variables (P17, P18, P20, and P21) relate to the satellite's power supply and indicate an absence of battery usage compared to a normal SIAs. The main power bus usage (P17) remained largely constant while the battery charge slowly increased (P18). The latter observation is explained by the nearly linear increase in battery charge current (P20). However, this linear charge current is anomalous given the solar array load current (P21), which fluctuates around a constant value. Notable, the second battery charge variable P19 is not considered a root cause candidate by our method. However, experts did also not find the related time series abnormal on its own.

Collectively, these observations suggest a failed image acquisition caused by a frozen processing unit or software bug that failed to initiate the imaging sequence.

Regarding the effectiveness of the proposed RCA method, it is worth noting that we did not flag the joint behavior of the attitude mode variables. This limitation is expected, as our generalized ordinal pattern analysis evaluates univariate statistics (normal vs. anomalous distributions) rather than multivariate dependencies. With the exception of P07–P09, the satellite engineer confirmed that no other variables appeared anomalous, which aligns with the expectation that a software issue would not manifest as physical sensor faults in other subsystems. Overall, the proposed root cause analysis successfully narrowed down the candidate variables in this example, streamlining the diagnostic and troubleshooting process.

3.4.2 Sensitivity analysis

Since our root cause analysis critically depends on the accuracy of the causal graph G_{com} , we conduct a sensitivity analysis to evaluate how potential structural errors affect our results. Following the procedure outlined in Sect. 2.2.4, we generated 20 perturbed graphs $G_{\text{com}}^{(s)}$, perturbing $\approx 5\%$ of the edges (i.e., 3 out of 68).

Applying our machine learning anomaly detection pipeline to these perturbed structures, we computed modified conditional z -scores $\{\tilde{z}_i^{(s)}\}_i$. We observe only partial agree-

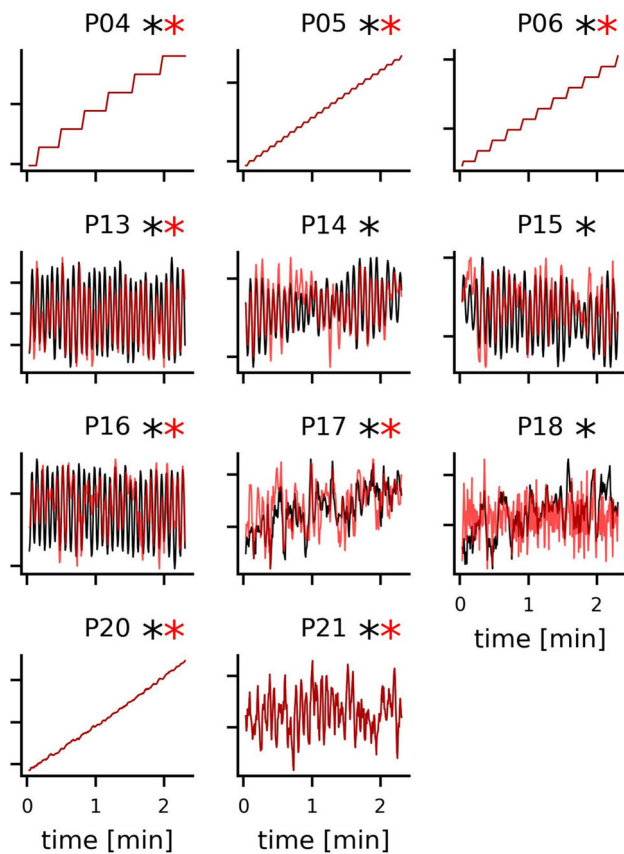


Fig. 7 Time series data of the exemplary anomalous SIA #95 for root cause analysis. Raw data (black) and conditional z -scores (red) of root cause candidate variables based on generalized ordinal pattern analysis. Black and red asterisks indicate root cause candidate variables identified with either raw data or conditional z -scores, respectively (cf. Fig. 6). For a comparison with normal behavior, see Fig. 3

ment between anomalies detected using the original versus perturbed scores ($F1 = 0.47$; precision = 0.44, recall = 0.49). Similarly, root cause candidates derived from generalized ordinal pattern analysis on the perturbed z -score time series show weak correspondence with those identified from the original analysis ($F1 = 0.37$; precision = 0.51, recall = 0.31).

These results underline that our approach is very sensitive to the correctness of G_{com} , which does not come as a surprise. Most advanced causal inference methods imply the correctness of the causal graph as a foundational assumption. To mitigate this risk, we rigorously constructed G_{com} using structural consistency checks and domain knowledge (cf. Sect. 2.2.1).

Although ground-truth validation of G_{com} is infeasible for this complex system, two lines of evidence support its reliability: (i) the marked improvement in anomaly detection performance when using conditional z -scores over raw time series data (cf. Table 3), and (ii) our expert-validated demonstration of root cause analysis on a satellite image acquisition

event. Together, these results suggest that G_{com} provides a structurally sound foundation for our analysis.

4 Summary and future work

This paper explores the use of anomaly detection and causal inference methods for satellite telemetry data. We demonstrated the investigation of uncharted anomalies and the attribution of root causes of anomalous behavior with a focus on explainable results that facilitate interpretation by satellite engineers.

For detecting anomalies in the telemetry data, we proposed an iterative process of anomaly exploration and selecting candidates using an ensemble approach of five different anomaly detection methods to find different types of anomalies. For each of these methods, we trained 100 models to increase the coverage of the anomaly space and relied on a majority vote to flag a satellite image acquisition (SIA) as anomalous. Following this approach, we identified 17 of 690 SIAs as anomalous.

To make the results explainable to the satellite engineers, we extended the anomaly detection framework with causal analysis. As the first step of this analysis, we employed the multi-data version of the time series causal discovery algorithm PCMC⁺, implemented within the Python package Tigramite in order to infer causal relationships between the system variables. We furthermore augment this data-driven causal analysis with expert knowledge, obtaining a common causal graph shared across all SIAs. However, in order for the inferred graph to meet the accuracy requirements of the subsequent analysis, we rely on expert knowledge and a substantial amount of time series data accumulated over multiple SIAs. To estimate the potential magnitude of errors induced by an incorrectly inferred causal structure, we employed a stability analysis inspired by surrogate methods for graphs [45] and found that our methods are indeed sensitive to the quality of the causal graph. While this highlights a limitation of causal discovery as tool for purely data-driven exploration of the satellite system, we employed the common causal graph to further investigate detected anomalies with root cause analysis methods and were able to achieve positive results in spite of potential issues with the causal graph.

Specifically, we explored an approach based on the work of Janzing et al. [38] and Budhathoki et al. [12] to detect of root cause candidate variables of outliers. Once anomalies were detected, our approach was capable of identifying candidates whose atypical behavior was not induced by causal parents. By comparing these candidates with the ones flagged by our anomaly detection investigation, we were able to either increase confidence in the identified root causes or attribute atypical behaviour to causal parents.

Our anomaly detection approach identified 17 outlier SIAs, of which 9 were confirmed by experts as potentially threatening to the system. While labeling the entire dataset by experts is impractical, the F1 score, precision, and recall evaluated on the small subset of SIAs investigated by experts suggest that anomaly detection on conditional z -score data (cf. Sect. 2.2.2) is more accurate at identifying non-benign anomalies when applied on the raw data.

Without access to the ground truth regarding the root causes, further assessment of our approach remains challenging. While we were able to assess our root cause analysis with the help of a domain expert in an example SIA, the statistical nature of our employed causal discovery algorithms and outlier detection methods make it impossible to rule out potential misattribution. Nevertheless, our results—especially, the confirmation of our example case—highlight the potential information gain from combining machine learning algorithms with root cause analysis in complex systems.

Overall, this paper contributes to the growing body of work on anomaly detection and causal inference in complex data sets (see, e.g., [12, 13, 16]), with a focus on developing explainable results that facilitate interpretation by satellite engineers. In a future work, we plan to integrate the results of the CausalAnomalies project into our novelty detection pipeline. Specifically, we aim to enhance ATHMoS to support multi-parameter capabilities, enabling a more nuanced and robust anomaly detection. Additionally, we intend to align the enabling causal inference framework with ATHMoS to incorporate temporal and spatial relationships between variables in the analysis of anomalies. By integrating these components, we expect to provide more explainable results and facilitate the interpretation of anomalies by satellite engineers.

Acknowledgements The authors appreciate the support of Steffen Zimmermann, head of the DLR RB-MIT department, as well as Jens Hartung Balan for his technical expertise, and the entire DLR RB-MBT team for their valuable input. We also thank Martin Rabel and Oana-Iuliana Popescu for their contributions to our previous work [22].

Author Contributions C.S., B.S., A.G., B.L., F.R., and T.R. wrote the main manuscript text. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets analyzed in this study are not publicly available due to confidentiality concerns. However, detailed descriptions of the data are provided in Sect. 3.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as

long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Schlag, L., Dauth, M., Braun, A.: The GSOC satellite telemetry analysis framework. In: Deutscher Luft- und Raumfahrtkongress 2019 (DLRK 2019) (2019). <https://elib.dlr.de/148456/>
- Schefels, C., Schlag, L., Del Moro, A., Helmsauer, K., Lesch, T., Göttfert, T.: Bringing a machine-learning based novelty detection software tool from research to production. In: Proceedings of the 17th International Conference on Space Operations (SpaceOps 2023) (2023). <https://elib.dlr.de/195314/>
- Helmsauer, K., Del Moro, A., Göttfert, T., Schefels, C., Schlag, L.: Ai-based novelty detection in space operations: three years of operational experience and progression at GSOC. In: 18th International Conference on Space Operations (SpaceOps2025), p. 0217 (2025). https://doi.org/10.82217/spaceops2025_217
- Del Moro, A., Lesch, T., Leidreiter, D., Helmsauer, K., Schefels, C., Göttfert, T.: Three (now four) years into operations for a satellite health monitoring system and its evolution. In: 18th International Conference on Space Operations (SpaceOps2025), p. 0272 (2025). https://doi.org/10.82217/spaceops2025_272
- O'Meara, C., Schlag, L., Faltenbacher, L., Wickler, M.: Atmos: Automated telemetry health monitoring system at GSOC using outlier detection and supervised machine learning. In: Proceedings of the 14th International Conference on Space Operations (SpaceOps 2016) (2016). <https://doi.org/10.2514/6.2016-2347>
- Schlag, L., O'Meara, C., Wickler, M.: Numerical analysis of automated anomaly detection algorithms for satellite telemetry. In: Proceedings of the 15th International Conference on Space Operations (SpaceOps 2018) (2018). <https://doi.org/10.2514/6.2018-2534>
- Schlag, L., Schefels, C., Helmsauer, K.: Applying machine learning to routine satellite ground segment operations by means of automated anomaly detection. In: Aerospace Europe Conference 2023 (EUCASS-CEAS 2023) (2023). <https://doi.org/10.13009/EUCASS2023-851>
- Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York (2000). <https://doi.org/10.1017/S0266466603004109>
- Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge (2000). <https://mitpress.mit.edu/9780262527927/>
- Imbens, G., Rubin, D.: Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge (2015). <https://doi.org/10.1017/CBO9781139025751>
- Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms, pp. 1214–1216. MIT Press, Cambridge (2017). <https://mitpress.mit.edu/9780262037310/>
- Budhathoki, K., Minorics, L., Blöbaum, P., Janzing, D.: Causal structure-based root cause analysis of outliers. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2357–2369. PMLR, Bal-

- timore (2022). <https://proceedings.mlr.press/v162/budhathoki22a.html>
13. Papageorgiou, K., Theodosiou, T., Rapti, A., Papageorgiou, E.I., Dimitriou, N., Tzovaras, D., Margetis, G.: A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing. *Front. Manuf. Technol.* **2**, 972712 (2022). <https://doi.org/10.3389/fmtec.2022.972712>
 14. Razouk, H., Liu, X.L., Kern, R.: Improving fmea comprehensibility via common-sense knowledge graph completion techniques. *IEEE Access* **11**, 127974–127986 (2023). <https://doi.org/10.1109/ACCESS.2023.3331585>
 15. Okazaki, S., Kaminishi, K., Fujii, T., Wang, Y., Ota, J.: Fault cause identification across manufacturing lines through ontology-guided and process-aware fmea graph learning with llms (2025). <https://doi.org/10.48550/arXiv.2510.15428>. [arXiv:2510.15428](https://arxiv.org/abs/2510.15428)
 16. Tappe, M., Moddemann, L., Steude, H., Hranisavljevic, N., Diedrich, A., Myschik, S., Geier, C., Creutzenberg, M., Grashorn, P., Hoppe, T., Ernst, H., Niggemann, O.: A supervised ai-based toolchain for anomaly detection, diagnosis, and reconfiguration for the life-support system of the columbus module of the iss. *CEAS Space J.* (2025). <https://doi.org/10.1007/s12567-025-00654-3>
 17. Liu, J., Li, F., Bao, J.: Afgnn: adaptive graph neural networks for causal failure reasoning in dfmea. *Res. Eng. Design* **37**(3), 27 (2026). <https://doi.org/10.1007/s00163-026-00490-4>
 18. Adelfio, A., Cutini, S., Longo, F., Germani, S., Lubrano, P., Crupi, R., Maldera, S.: Advances on Tslies: time series anomalies, an efficient machine learning based anomaly detection framework. *Astron. Comput.* **56**, 101113 (2026). <https://doi.org/10.1016/j.ascom.2026.101113>
 19. Ares-Robledo, F., Rifà-Pous, H., Clarisó, R.: Graph neural networks for anomaly detection: a systematic review of dynamic temporal approaches. *Artif. Intell. Rev.* **59**, 129 (2026). <https://doi.org/10.1007/s10462-026-11532-7>
 20. Nguyen, H.H., Nguyen, C.N., Dao, X.T., Duong, Q.T., Kim, D.P.T., Pham, M.-T.: Variational autoencoder for anomaly detection: a comparative study (2024). <https://doi.org/10.48550/arXiv.2408.13561>. [arXiv:2408.13561](https://arxiv.org/abs/2408.13561)
 21. Shi, J., Tang, M., Tan, L.: Backbone agnostic pareto evidential networks for trustworthy fault diagnosis and out of distribution detection. *Sci. Rep.* **16**, 10096 (2026). <https://doi.org/10.1038/s41598-026-40463-0>
 22. Schefels, C., Ben Salem, B., Gerhardus, A., Helmsauer, K., Lambert, B., Niebling, J., Popescu, O.-I., Rabel, M., Rewicki, F., Schlag, L.: Explaining satellite anomalies-causal inference for space operations. In: 18th International Conference on Space Operations (SpaceOps 2025) (2025). https://doi.org/10.82217/spaceops2025_338
 23. Rewicki, F., Denzler, J., Niebling, J.: Is it worth it? Comparing six deep and classical methods for unsupervised anomaly detection in time series. *Appl. Sci.* (2023). <https://doi.org/10.3390/app13031778>
 24. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: 8th IEEE International Conference on Data Mining, pp. 413–422 (2008). <https://doi.org/10.1109/ICDM.2008.17>
 25. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. *Neural Comput.* **13**(7), 1443–1471 (2001). <https://doi.org/10.1162/089976601750264965>
 26. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. *Pattern Recogn. Lett.* **24**(9), 1641–1650 (2003). [https://doi.org/10.1016/S0167-8655\(03\)00003-5](https://doi.org/10.1016/S0167-8655(03)00003-5)
 27. Dai, E., Chen, J.: Graph-augmented normalizing flows for anomaly detection of multiple time series, virtual (2022). <https://iclr.cc/virtual/2022/spotlight/6771>
 28. Barz, B., Rodner, E., Garcia, Y.G., Denzler, J.: Detecting regions of maximal divergence for spatio-temporal anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(5), 1088–1101 (2019). <https://doi.org/10.1109/TPAMI.2018.2823766>
 29. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. *Front. Genet.* **10**, 524 (2019). <https://doi.org/10.3389/fgene.2019.00524>
 30. Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., Runge, J.: Discovering causal relations and equations from data. *Phys. Rep.* **1044**, 1–68 (2023). <https://doi.org/10.1016/j.physrep.2023.10.005>
 31. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* **9**(1), 62–72 (1991). <https://doi.org/10.1177/089443939100900106>
 32. Saranya Ganesh, S., Beucler, T., Tam, F.I.-H., Gomez, M.S., Runge, J., Gerhardus, A.: Selecting robust features for machine-learning applications using multitask causal discovery. *Environ. Data Sci.* **2**, 27 (2023). <https://doi.org/10.1017/eds.2023.21>
 33. Runge, J.: Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In: Peters, J., Sontag, D. (eds.) 36th Conference on Uncertainty in Artificial Intelligence (UAI). *Proceedings of Machine Learning Research*, vol. 124, pp. 1388–1397. PMLR, virtual (2020). <https://proceedings.mlr.press/v124/runge20a.html>
 34. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**(11), 4996 (2019). <https://doi.org/10.1126/sciadv.aau4996>
 35. Tsagris, M., Borboudakis, G., Lagani, V., Tsamardinos, I.: Constraint-based causal discovery with mixed data. *Int. J. Data Sci. Anal.* **6**, 19–30 (2018). <https://doi.org/10.1007/s41060-018-0097-y>
 36. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65**, 31–78 (2006). <https://doi.org/10.1007/s10994-006-6889-7>
 37. Alinezhad, H.S., Roohi, M.H., Chen, T.: A review of alarm root cause analysis in process industries: common methods, recent research status and challenges. *Chem. Eng. Res. Des.* **188**, 846–860 (2022). <https://doi.org/10.1016/j.cherd.2022.10.041>
 38. Janzing, D., Budhathoki, K., Minorics, L., Blöbaum, P.: Causal Structure Based Root Cause Analysis of Outliers (2019). [arXiv:1912.02724](https://arxiv.org/abs/1912.02724)
 39. Bollen, K.A.: *Structural Equations with Latent Variables*. Wiley, New York (1989). <https://doi.org/10.1002/9781118619179>
 40. Bishop, C.M., Nasrabadi, N.M.: *Pattern Recognition and Machine Learning*, vol. 4. Springer, New York (2006). <https://link.springer.com/book/9780387310732>
 41. Bandt, C., Pompe, B.: Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett.* **88**(17), 174102 (2002). <https://doi.org/10.1103/PhysRevLett.88.174102>
 42. Silbernagel, A., Schnurr, A.: An overview of various applications of ordinal patterns in data analysis and mathematical statistics. In: International Conference on Electrical, Computer and Energy Technologies (ICECET 2024), pp. 1–5. IEEE, New York (2024). <https://doi.org/10.1109/ICECET61485.2024.10698110>
 43. Schnurr, A., Fischer, S.: Generalized ordinal patterns allowing for ties and their applications in hydrology. *Comput. Stat. Data Anal.* **171**, 107472 (2022). <https://doi.org/10.1016/j.csda.2022.107472>
 44. Agresti, A.: *Categorical Data Analysis*. Wiley, Hoboken (2013). <https://doi.org/10.1002/0471249688>
 45. Ansmann, G., Lehnertz, K.: Constrained randomization of weighted networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **84**, 026103 (2011). <https://doi.org/10.1103/PhysRevE.84.026103>