SOLENIX

Controlled Anomalies Time Series (CATS) Dataset

Dataset Description Document – Version 2

The Controlled Anomalies Time Series (CATS) Dataset consists of commands, external stimuli, and telemetry readings of a simulated complex dynamical system with 200 injected anomalies.

This document refers to version 2 of the CATS dataset. See the "**Change Log**" section below for a detailed description of the improvements over version 1 of the dataset.

Dataset Description

The CATS Dataset exhibits a set of desirable properties that make it very suitable for benchmarking **Anomaly Detection Algorithms in Multivariate Time Series**[1]:

- **Multivariate (17 variables)** including sensors reading and control signals. It simulates the operational behaviour of an arbitrary complex system including:
    - **4 Deliberate Actuations / Control Commands sent by a simulated operator / controller**, for instance, commands of an operator to turn ON/OFF some equipment.
    - **3 Environmental Stimuli / External Forces** acting on the system and affecting its behaviour, for instance, the wind affecting the orientation of a large ground antenna.
    - **10 Telemetry Readings** representing the observable states of the complex system by means of sensors, for instance, a position, a temperature, a pressure, a voltage, current, humidity, velocity, acceleration, etc.
- **5 million timestamps**. Sensors readings are at 1Hz sampling frequency.
    - **1 million nominal** observations (the first 1 million datapoints). This is suitable to start learning the "normal" behaviour.
    - **4 million** observations that include both **nominal and anomalous segments**. This is suitable to evaluate both semi-supervised approaches (novelty detection) as well as unsupervised approaches (outlier detection).
- **200 anomalous segments.** One anomalous segment may contain several successive anomalous observations / timestamps. Only the last 4 million observations contain anomalous segments.
    - **Contamination level of 0.038.** This means about 3.8% of the observations (rows) are anomalous.
- **Different types of anomalies** to understand what anomaly types can be detected by different approaches. The categories are available in the dataset and in the metadata.
- **Fine control over ground truth.** As this is a simulated system with deliberate anomaly injection, the start and end time of the anomalous behaviour is known very precisely. In contrast to real world datasets, there is no risk that the ground truth contains mislabelled segments which is often the case for real data.

[1] Example Benchmark of Anomaly Detection in Time Series: "Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly Detection in Time Series: A Comprehensive Evaluation. PVLDB, 15(9): 1779 - 1797, 2022. doi:10.14778/3538598.3538602"

- **Suitable for root cause analysis.** In addition to the anomaly category, the time series channel in which the anomaly first developed itself is recorded and made available as part of the metadata. This can be useful to evaluate the performance of algorithm to trace back anomalies to the right root cause channel.

- **Affected channels.** In addition to the knowledge of the root cause channel in which the anomaly first developed itself, we provide information of channels possibly affected by the anomaly. This can also be useful to evaluate the explainability of anomaly detection systems which may point out to the anomalous channels (root cause and affected).

- **Obvious anomalies.** The simulated anomalies have been designed to be "easy" to be detected for human eyes (i.e., there are very large spikes or oscillations), hence also detectable for most algorithms. It makes this synthetic dataset useful for screening tasks (i.e., to eliminate algorithms that are not capable to detect those obvious anomalies). However, during our initial experiments, the dataset turned out to be challenging enough even for state-of-the-art anomaly detection approaches, making it suitable also for regular benchmark studies.

- **Context provided.** Some variables can only be considered anomalous in relation to other behaviours. A typical example consists of a light and switch pair. The light being either on or off is nominal, the same goes for the switch, but having the switch on and the light off shall be considered anomalous. In the CATS dataset, users can choose (or not) to use the available context, and external stimuli, to test the usefulness of the context for detecting anomalies in this simulation.

- **Pure signal ideal for robustness-to-noise analysis.** The simulated signals are provided without noise: while this may seem unrealistic at first, it is an advantage since users of the dataset can decide to add on top of the provided series any type of noise and choose an amplitude. This makes it well suited to test how sensitive and robust detection algorithms are against various levels of noise.

- **No missing data.** You can drop whatever data you want to assess the impact of missing values on your detector with respect to a clean baseline.

Figure 1 below depicts the name of each time series variable and its relation with respect to the simulated operator, environmental stimuli, and operated system.

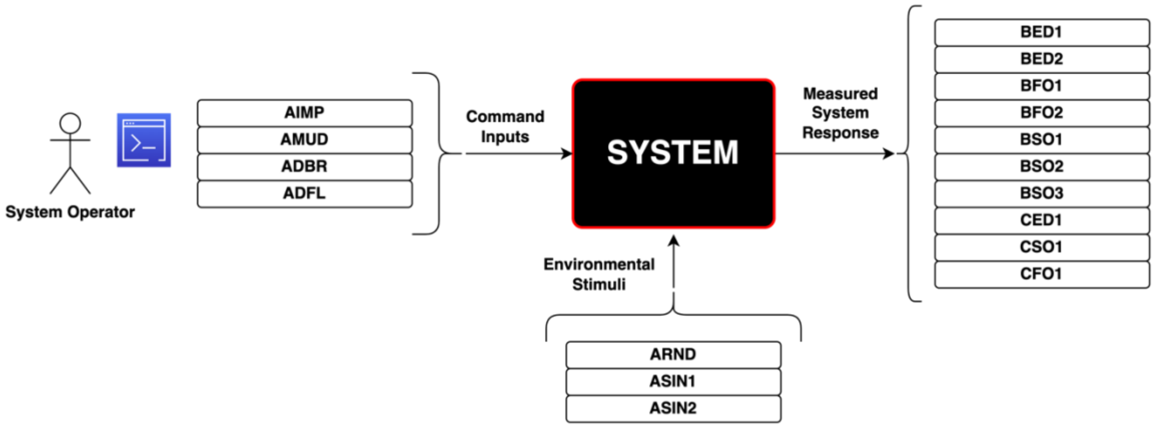

Figure 1 - Input/output relations with time series signal and the simulated system

Change Log

Version 2

- **Metadata:** we include a metadata.csv with information about:
    - Anomaly categories
    - Root cause channel (signal in which the anomaly is first visible)
    - Affected channel (signal in which the anomaly might propagate) through coupled system dynamics
- **Removal of anomaly overlaps:** version 1 contained anomalies which overlapped with each other resulting in only 190 distinct anomalous segments. Now, there are no more anomaly overlaps.
- **Two data files:** CSV and parquet for convenience.

License

About Solenix

Solenix is an international company providing software engineering, consulting services and software products for the space market. Solenix is a dynamic company that brings innovative technologies and concepts to the aerospace market, keeping up to date with technical advancements and actively promoting spin-in and spin-out technology activities. We combine modern solutions which complement conventional practices. We aspire to achieve maximum customer satisfaction by fostering collaboration, constructivism, and flexibility.