# Coupling Variable Selection and Anomaly Detection: Record-Based Approach

Michel Kamel, Anis Hoayek, Mireille Batton-Hubert

# Coupling Variable Selection and Anomaly Detection: Record-Based Approach

Michel Kamel[1*], Anis Hoayek[1], Mireille Batton-Hubert[1]

Abstract

The proliferation of interconnected devices is rapidly expanding globally, and, as a result, telecommunication operators are responsible for managing intricate and expansive networks. Consequently, there is a need for advanced and efficient systems to aid network engineers in maintaining these networks. Devices, which can also be referred to as network elements, continuously transmit essential performance data known as key performance indicators. By utilizing data derived from these metrics and implementing intelligent anomaly detection models, the devices can assist in determining the optimal production maintenance schedule for the network. As anomaly detection models deal with extreme events, this study proposes a method of reducing dimensions by focusing on the behavior of the tails of underlying variables, rather than the entire distribution. In addition to that, an anomaly scoring system, also based on records theory, is proposed, which has several advantages over current state-of-the-art models. The effectiveness of this approach is demonstrated by implementing it on a real-world dataset.

## I. Introduction

Identifying anomalies in a time series refers to identifying observations that differ from the typical pattern of the other observations. Such anomalies are uncommon and significant as they can have an impact on the underlying system that generates the time series. It is crucial to quickly and accurately detect these abnormal behaviors to ensure the proper functioning of the upstream system (Chandola et al., 2009). Anomaly detection is a research topic that is encountered in various fields such as industry (Zhou et al., 2020), cybersecurity (Rashid et al., 2022), healthcare (Sabic et al., 2021), environment (Vangipuram et al., 2020), and telecommunication (Kamel et al., 2023; Ali et al., 2020).

Various research areas, including machine learning (ML) (Alvi et al., 2022), statistical learning (Sha et al., 2015), game theory (Huang et al., 2019), and graph theory (Akoglu et al., 2015) have contributed to the development of current state-of-the-art anomaly detection algorithms and models.

The literature has addressed several questions and challenges related to anomaly detection. One such challenge is the ability to generate abnormality scores in an unsupervised context, where auto-encoder (AE) neural network models are the most widely used approach. Another challenge is detecting abnormal behavior without making any assumptions about the probabilistic distribution of the underlying random variables, which is addressed by using random-forest–based models. However, there are still many challenges that cannot be tackled by traditional anomaly detection models and require innovative approaches.

[1] *Ecole des Mines de Saint-Etienne Univ. Clermont Auvergne, CNRS UMR 6158 LIMOS Mathematics and Industrial Engineering*, Saint-Etienne, France, michel.kamel@emse.fr

[*] Corresponding author.

This paper aims to address several research gaps in anomaly detection. First, it proposes a dimension reduction method that is adapted to abnormal and extreme events for dealing with large datasets. This method uses records theory (see Sections II and III for details) which is a branch of extreme value theory to develop a variable selection methodology that focuses on the behavior of the tails of the underlying variables rather than the whole distribution, as in classical dimension reduction methods like principal component analysis and AE. Second, the paper uses records theory to propose an abnormality scoring system that can be used in one or multidimensional datasets. This system generates a density distribution of the scores and uses a grid search process to minimize classification errors and set a threshold value for the underlying variables above which an observation is considered abnormal. This threshold can be communicated to subject matter experts (SMEs), and, to the best of the authors' knowledge, this is the first algorithm to propose an anomaly threshold for each considered variable. Third, the proposed anomaly detection algorithm is adapted to online learning modes and is optimized in terms of computational complexity, despite being coupled with a variable selection method. Finally, the proposed approach provides initial information about the root cause of a detected anomaly.

In summary, the main objective of this paper is to develop a comprehensive methodology that enables users to detect anomalies in time series data while addressing the challenges associated with this task.

The proposed method is primarily designed to detect abnormal behavior in different elements of a telecommunication network. These network elements generate numerous key performance indicators (KPIs), and the sheer number of features that describe the performance of the different services provided is enormous, making manual analysis of these observations challenging, if not impossible. In addition, the ability to identify irregularities in real time with minimal delays requires the utilization of sophisticated correlation analysis and extensive data mining techniques to reveal concealed patterns and associations within the generated data.

The rest of this paper is structured as follows. Section II provides an introduction to records theory. Section III presents the mathematical formalization of the most popular models in records theory and how they are adapted to the current context. Section IV describes the use of records theory for variable selection. Section V shows how records theory is used to generate abnormality scores in one and multidimensional datasets. Section VI demonstrates some real-world applications. Finally, Section VII concludes the paper.

## II. Records, an Introduction

The study of records in a time series as a field of extreme value theory can be traced back to Chandler's work in 1952. Since then, there have been numerous developments in this field, including the works of Arnold, Nevzorov, and their collaborators during the 1980s and 1990s. Initially, researchers focused on the classic case in records theory, which assumes that the random variables (RV) are independent and identically distributed (IID). However, this case did not fully capture the complexity of multiple datasets, so researchers began to explore cases where the observations are independent but not identically distributed. Eventually, they even considered the most general case where neither the independence assumption nor the assumption of identical distribution holds.

Data are found in the form of records across various fields that use statistics, such as sports (Yang, 1975), climate change (Wergen and Krug, 2010; Wergen, 2013), risk assessment of diseases (Khraibani et al., 2015), financial markets (Hoayek et al., 2018), and satellite imagery (Jabbour et al., 2021).

It is worth noting that there is a greater interest in records when they are the only available values in a particular time series. Since records are a part of popular culture, they are usually kept in easily accessible places, such as the Guinness World Records.
In simpler terms, a record is a result in a given series of events that exceeds anything seen before. Therefore, a new record is always something remarkable and attracts attention, whether it is associated with positive or negative news.
Our research applies records theory to solve two challenges related to anomaly detection in an industrial context. The first challenge is to reduce the dimensionality when dealing with a large number of time series to detect abnormal behavior. The second challenge is to develop an innovative ML model that efficiently and accurately detects anomalies using the principles of records theory, which aims to model extreme values.

## III. Mathematical Formalization

We start by considering the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Here, $X$ denotes a real RV with a cumulative distribution function (CDF) $F(\cdot)$ and a density function $f(\cdot)$. We assume that the space $(\Omega, \mathcal{F}, \mathbb{P})$ has good properties to define an infinite sequence $\{X_t, t \geq 1\}$ of IID RVs, which are independent copies of $X$. When the index $t$ represents time, then we are dealing with a time series having an IID underlying distribution. An observation $X_t$ is considered an upper record at time $t$ if it is higher than all previous observations, that is, $X_t > \max(X_1, \dots, X_{t-1})$. In this paper, we focus on upper records, but lower records can be defined similarly, by multiplying the time series by "–1". As time progresses, another important sequence of RVs can be defined: the sequence of record values $\{R_n, n \geq 1\}$ and the sequence of occurrence time of records $\{L_n, n \geq 1\}$. In other words, $L_n$ is the occurrence time of the $n^{th}$ record, which is $R_n = X_{L_n}$.
In most applications of records theory, the available data consists of a sequence of pairs $\{(R_n, L_n), n = 1, \dots, N_T\}$, where $T$ represents the current time (i.e., length of the time series) and $N_T$ is the total number of records in $\{X_t, t = 1, \dots, T\}$.
In addition to the previously defined sequences, one can also define the sequence of record indicators $\{\delta_t, t = 1, \dots, T\}$, where:

$$\delta_t = \begin{cases} 1 \text{ if } X_t \text{ is a record} \\ 0 \quad \text{otherwise} \end{cases}. \qquad (1)$$

Note that $\delta_1 = 1$ because the first observation is always a record, which is called a trivial record. We will demonstrate later, in this section, that based solely on the sequence of record indicators, we can extract significant information about the overall behavior of the records in a time series. Note that it is straightforward to remark that:

$$N_T = \sum_{t=1}^{T} \delta_t.$$

The stochastic properties of sequences of record values have been widely studied in the case where $X_t$ are IID RVs (Arnold et al., 2011; Nevzorov, 2001). Many of these properties are distribution-free, meaning that they do not depend on the choice of the underlying distribution of the observations. The most important results in the IID context are:
First, $\forall t \geq 1$, $\delta_t$ are mutually independent and follow a Bernoulli distribution with parameter $P_t = \frac{1}{t}$, which is called the record rate at time $t$. In other words, $\mathbb{P}[\delta_t = 1] = \frac{1}{t}$, which is the probability of observing a record at time $t$, and $\mathbb{E}[\delta_t] = \frac{1}{t}$. It is worth noting that:

$$\lim_{t \to +\infty} P_t = 0. \qquad (2)$$

Therefore, one can conclude that records are more likely to appear among the first observations. In addition, the expected number of records until time $T$ is given by:

$$\mathbb{E}[N_T] = \sum_{t=1}^{T} \mathbb{E}[\delta_t],$$
$$= \sum_{t=1}^{T} \frac{1}{t}. \qquad (3)$$

Moreover, Arnold et al. (2011) found that records tend to become more spread out over time a $t$ or $n$ increases. However, this was not always the case in practice. For example, advancements in technology are causing sports records to occur more frequently than what is expected under the IID assumption. As a result, more complex models have been developed to better predict records beyond the classical IID case. These models can be grouped into two families based on their level of complexity, which we will discuss in the next two subsections.

A. Independent but Not Identically Distributed Observations

First, consider the case where underlying observations are independent but not identically distributed. In this context, two common models are used:

- Linear drift model (LDM), introduced by Ballerini and Resnick in 1985, formalized by:
$$X_t = Y_t + \theta t, \qquad (4)$$
where $Y_t, t \geq 1$ is a sequence of IID RVs and $\theta > 0$ is a parameter that needs to be estimated.

- Yang record model, initially introduced by Yang (1975) and later developed by Nevzorov (1988). This model is considered more suitable for the independent but not identically distributed context and in most cases, it is more generalized than the LDM. The Yang model can be represented by the following formula:
$$X_t \sim F(\cdot)^{\rho_t}, \qquad (5)$$
where $\rho_t (t \geq 1)$ are real constants $\geq 1$ and $F(\cdot)$ is a CDF of a particular underlying distribution. In this paper, we will focus on a specific parametrization of the Yang model, in which $\rho_t = \gamma^t$, with $\gamma$ being a parameter that needs to be estimated and is $\geq 1$. This formalization is interesting because it has the structure of a proportional hazard model, which is commonly used in survival analysis to model various datasets (Hoayek et al., 2017). In addition, each $X_t$ represents the maximum value obtained from $\rho_t$ observations that are generated simultaneously and independently at time $t$ from the same underlying RV $Y$ of CDF $F(\cdot)$. Then,
$$X_t = \max(y_1, y_2, \dots, y_{\rho_t}). \qquad (6)$$
Based on the fact that the underlying RV $Y$ is independent, it can be demonstrated that the record rate at time $t$ is expressed as follows:
$$P_t = \mathbb{P}[\delta_t = 1] = \frac{\rho_t}{\sum_{k=1}^{t} \rho_k},$$
$$P_t = \frac{\gamma^t}{\sum_{k=1}^{t} \gamma^k} = \frac{\gamma^t(\gamma - 1)}{\gamma(\gamma^t - 1)}. \qquad (7)$$
In this case $P_t$ will be denoted as $P_t(\gamma)$.
Thus,
$$\lim_{t \to +\infty} P_t(\gamma) = \lim_{t \to +\infty} \frac{(\gamma - 1)}{\gamma\left(1 - 1/\gamma^t\right)} = \frac{\gamma - 1}{\gamma}. \qquad (8)$$

Therefore, in the Yang model, the probability of having new records in the long term does not decrease. As a result, a time series exhibiting this type of behavior can be considered more volatile and unstable compared to the classical IID case.

Despite its usefulness in various applications, the Yang model cannot be utilized in practice without first estimating the parameter $\gamma$. To do this, Hoayek et al. (2017) proposed an estimation method based on maximizing the following Log-Likelihood function that was constructed using solely the observed sequence of indicators:

$$\mathrm{Log}\, L(\gamma) = \mathrm{Log}\, \mathbb{P}[\delta_1, \dots, \delta_T; \gamma]. \qquad (10)$$

Then, by solving,

$$\frac{\mathrm{d}\, \mathrm{Log}\, L(\gamma)}{\mathrm{d}\gamma} = 0, \qquad (11)$$

we get our estimator which is denoted by $\hat{\gamma}$. In addition, also based on the work of Hoayek et al. (2017), one can show the asymptotic behavior of $\hat{\gamma}$ which is also distribution-free:

$$\frac{(\hat{\gamma} - \gamma)}{\sqrt{I_T^{-1}(\gamma)}} \longrightarrow N(0,1), \qquad (12)$$

Here, $I_T^{-1}(\gamma)$ represents the Fisher information associated with the previous likelihood. Therefore, by understanding the asymptotic behavior of our estimator, we can conduct further inferential analysis such as constructing confidence intervals for a given asymptotic risk of error level $\alpha$.

Additionally, in the same context, Nevzorov (1988) demonstrated that record indicators are mutually independent, regardless of the choice of the underlying distribution $Y$. Thus, it can be concluded that the stochastic process $\{\delta_t\}_{t\geq 1}$ is a Bernoulli process with parameter $P_t$. Using this property, we can obtain the expression of the expected value and variance of the number of records:

$$\mathbb{E}[N_T] = \sum_{t=1}^{T} \mathbb{E}[\delta_t] = \sum_{t=1}^{T} P_t, \qquad (13)$$

$$\mathbb{V}[N_T] = \sum_{t=1}^{T} \mathbb{V}[\delta_t] = \sum_{t=1}^{T} P_t\,(1 - P_t) = \mathbb{E}[N_T] - \sum_{t=1}^{T} {P_t}^2. \qquad (14)$$

B. Dependent and Not Identically Distributed Observations

Another level of complexity arises when we consider the scenario where underlying observations are dependent and not identically distributed. In this context, the most prevalent record model is the discrete-time random walk model (DTRW) introduced by Majumdar and Ziff (2008). The underlying observations in this model can be formalized as follows:

$$X_t = X_{t-1} + \eta_t, \qquad (15)$$

where the increments $\eta_t$ are drawn from a continuous distribution in an IID way.

In the context of DTRW, the record rate at time $t$ can be expressed as:

$$P_t = \mathbb{P}[\delta_t = 1] = \binom{2t}{t} 2^{-2t}. \qquad (16)$$

Majumdar and Ziff (2008) demonstrated that $P_t$ asymptotically approaches zero in the DTRW model, though at a slower rate than it does for the IID case. Therefore, it can be concluded that in terms of long-term record probability, the DTRW model lies somewhere between the classical IID model and the Yang model. Additionally, it is worth noting that the majority of the results on DTRW are distribution-free.

Figure I provides a comprehensive overview of the behavior of record rates for different record models. Note that in Figure I, without any loss of generality, the parameter $\gamma$ of the Yang model is assumed to be equal to 1.2.

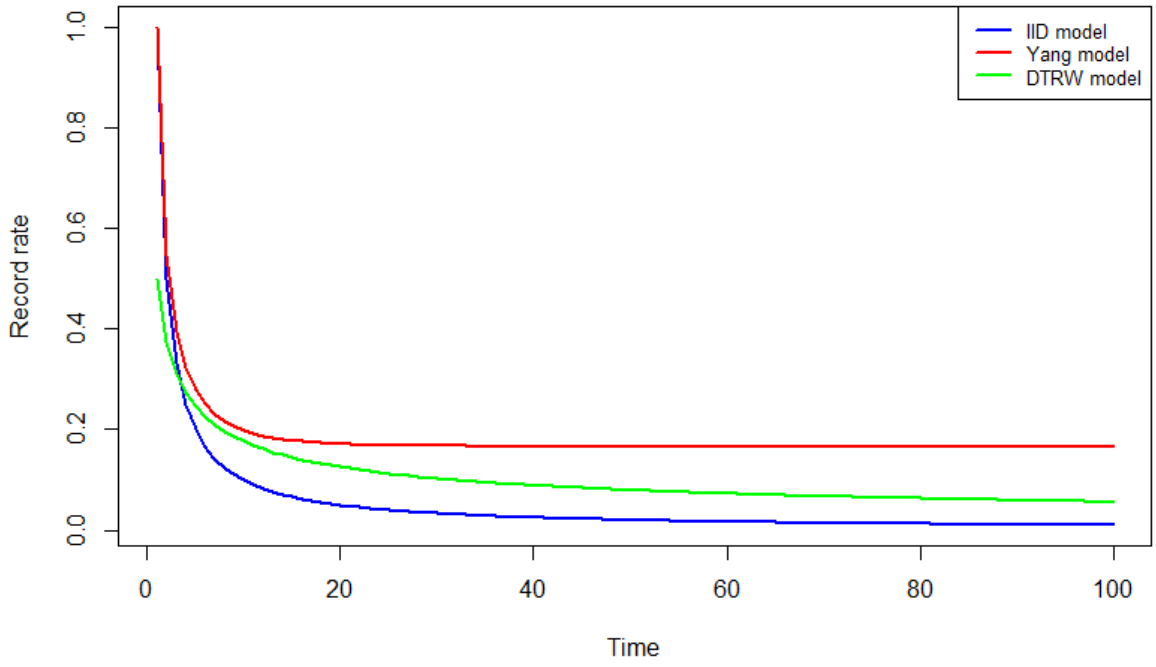



Figure I: Record rates for different record models for a time series of length 100.

C. Record Model Selection

Selecting the appropriate model to explain the record behavior of a time series involves performing a sequence of statistical tests.

Before considering non-IID models, we begin by testing whether the underlying observations of the time series in question are generated from an IID sequence. We do this by considering the null hypothesis:
$H_0$: Records are generated from an IID underlying sequence of observations.
To perform this first test, we use the fact that under $H_0$, Arnold et al. (2011) showed that:

$$\mathcal{N}_T = \frac{N_T - \log T}{\sqrt{\log T}} \longrightarrow \text{Standard Gaussian Distribution } N(0,1). \qquad (17)$$

Therefore, $\mathcal{N}_T$ can be viewed as the statistic used in the test. In practice, if $\mathcal{N}_T > q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of $N(0,1)$, then the IID model is rejected, and we should consider one of the models outside the classical case.

The next step is to establish a statistical test for the Yang model. Assuming the Yang model hypothesis, we create an RV referred to as the inter-record time (i.e., the time between two consecutive records), which is defined as:

$$\Delta_{L_n} = L_{n+1} - L_n, n \geq 1. \qquad (18)$$

Hoayek et al. (2017) showed that in a Yang model $\Delta_{L_n}$ follow a geometric distribution asymptotically. Therefore, we can use this result to construct a goodness of fit test for the Yang model. The null hypothesis for this test is that the inter-record time observations fit a geometric

distribution. To conduct this test, we can adapt Pearson's chi-square test to the context of record models (for details see Hoayek et al., 2017). If Pearson's test rejects the geometric distribution, it is not appropriate to use the Yang model, and we should consider moving to a higher level of complexity where observations are dependent and not identically distributed. A common method to test the dependency between underlying observations is to use the Ljung–Box test (1978).

## IV. Variable Selection Based on Record Behavior

We will apply the methodology outlined in the previous section to create a variable selection tool for detecting anomalies. Most anomaly detection algorithms are designed to identify abnormal behavior, and users often aim to avoid the problem of dimensionality that can lead to increased computational costs, especially during the training phase.
To address this issue, various classical solutions have been proposed, such as linear and non-linear dimension reduction methods like principal component analysis and AEs, as well as variable selection methods like genetic algorithms. However, all of these methods consider the entire multidimensional distribution behavior of the underlying variables to determine how dimension reduction should be performed. This approach may not be suitable for certain application contexts, particularly when the focus is on the tails distribution behavior of the variables, as in the fields of anomaly detection and extreme event detection. Therefore, the proposed method is innovative and specifically adapted for the anomaly detection case. The method primarily focuses on the behavior of extreme events, especially upper records, to determine which variables should be selected. By prioritizing the features that are critical during anomaly detection, the dimension of the decision space is reduced, and the application of any algorithm becomes faster and less computationally complex. This is particularly important for online detection purposes.

In practice, we will use a collection of time series, known as KPIs, to evaluate the quality of services provided by a virtual telecommunication network. These KPIs will be used to identify abnormal behavior for each element of the network.
However, before applying any of the anomaly detection algorithms, we will assess each KPI separately and assign a priority level to each of them based on the following rules:

a. KPIs with high priority: when record behavior related to the underlying distribution of the KPI follows the Yang model. In this case, the probability of observing a new record on a long-term basis is constant, and extreme abnormal behavior is always likely to occur at any time. Therefore, such KPIs are considered risky in the context of anomaly detection.
b. KPIs with medium priority: when the DTRW model is accepted as a description of record behavior. In this case, record rates converge to zero in the long term, but at a slower rate than in the classical record IID model (see Figure I). Such records are considered to have medium risk and may have a significant impact on abnormal behavior.
c. KPIs with low priority: when records fit the classical IDD case. In such cases, record rates converge rapidly to zero, and abnormal behavior is observed very rarely on a long-term basis. Such KPIs are considered to have low risk and can be removed from later global abnormal behavior analysis.

To track the changes in the behavior of KPIs over time, we will apply the priority classification rules on sliding windows of fixed length $'k'$ with a step size $'s'$. This will allow us to monitor the KPIs' behavior over time and make any necessary adjustments to their priority levels. After

assessing each window, a final decision on the priority level of the KPI will be made by aggregating the results of all the windows, using a rule determined by SMEs.
In practical terms, the entire time series is examined for each KPI and then broken down into sliding windows. Using the information gathered from each window, we carry out the following steps:

- Step 0: If the KPI shows concerning high values, do not change the time series observations $X_t^w, t = 1, \dots, N = k$, where $X_t^w$ denotes the observation of the considered KPI at time $t$ in window $w$. Otherwise, transform the time series by considering $-X_t^w$ instead of $X_t^w$. In both cases, focus on the upper records for analysis.
- Step 1: From the time series obtained at the end of Step 0, extract record observations $(R_n, L_n)$ and calculate the values of record indicators RV $\delta_t$ and the number of records $N_T$.
- Step 2: Test for an IID behavior based on the statistic of Eq. (17). If the behavior is classical, assign a low priority to the window. Otherwise, proceed to Step 3.
- Step 3: Calculate the values of the inter-record times RV $\Delta_{L_n}$ and use them to perform the goodness of fit test for the Yang model. If the Yang model is reasonable, assign a high priority to the considered window. Otherwise, consider that we are in the context of the DTRW model and assign a medium priority to the considered window.
- Step 5: Repeat Steps 1 to 3 for all sliding windows and assign priority decisions for each window.
- Step 6: Aggregate the results for all windows using a rule established by SMEs and assign the resultant priority to the corresponding KPI. For example, consider the highest priority assigned across all windows as the KPI priority.

## V. Anomaly Scoring System Based on Records Distribution

### A. One Dimensional Abnormality Score

It is crucial to create a scoring system based on records to detect anomalies in a random variable that exhibits extreme behavior and abnormal events.
Suppose $\{X_t, t \geq 1\}$ is a time series that represents the behavior of a specific KPI over time with real values. $\forall t \geq 1$, we denote:

$$\Lambda_t = \left\{R_n, n \geq 1 \text{ such that } R_n \text{ is the } n^{th} \text{ record of the series } \{X_i, 1 \leq i < t\}\right\},$$
$$\Lambda_t^* = \{R_n \in \Lambda_t \text{ such that } R_n \geq X_t\},$$
$$\mathcal{D}_t = \{R_n - X_t, \text{such that } R_n \in \Lambda_t^*\},$$
$$\overline{\mathcal{D}_t} = \text{Arithmetic average of the elements of } \mathcal{D}_t.$$

Now for each observation $X_t, t \geq 1$ the corresponding abnormality score is given by:

$$\mathcal{S}_t = \begin{cases} 1, & \text{if } X_t \text{ is a record} \\ \dfrac{1}{\left(1 + \dfrac{\text{Card}\,\Lambda_t^*}{\text{Card}\,\Lambda_t}\right)} \times \left(\dfrac{1}{1 + \overline{\mathcal{D}_t}}\right), & \text{Otherwise} \end{cases}.$$

Where $\text{Card}(\cdot)$ gives the number of elements in a given set.
Assuming that the time series $\{X_t, t \geq 1\}$ has been standardized to have values between 0 and 1, and transformed so that high values indicate abnormal behavior, the $\mathcal{S}_t$ will fall between $\frac{1}{4}$

and 1. This score is calculated based on upper records only. Whenever $X_t$ reaches its maximum (i.e, $X_t = 1$), it is considered a new record.
On the other hand, when $\mathcal{S}_t$ is closer to 1, it indicates a higher risk of abnormal behavior. Each component of $\mathcal{S}_t$ focuses on an aspect of abnormality in the underlying time series based on records:

- $\frac{1}{\left(1+\frac{\text{Card}\,\Lambda_t^*}{\text{Card}\,\Lambda_t}\right)}$: This component is closer to 1 when almost all the records taking place before $t$ are lower than $X_t$. Therefore, in this case, even if $X_t$ is not a record, it has an impact that is comparable to the majority of the previously detected records and should be highlighted as a potential anomaly.
- $\left(\frac{1}{1+\overline{\mathcal{D}_t}}\right)$: This component has a complementary role to the previous one. Here, we are computing the average distance between the observation $X_t$ and all previously detected records with a value higher than $X_t$ (elements of $\Lambda_t^*$). Thus, for this component, we obtain a value close to 1 when the value of $X_t$ is close to the records of the set $\Lambda_t^*$ which is also a scenario that should be highlighted in the process of detecting potential abnormal behavior.

While not an exhaustive list, the proposed record-based scoring system offers several advantages over classical anomaly detection models:

1. Unlike popular anomaly detection ML models, there is no risk of overfitting because there is no classical training/testing phase in the proposed algorithm. Additionally, the algorithm is designed to function as an online anomaly score system, generating a score for each new arrival.
2. The algorithm is distribution-free, meaning there is no need to make assumptions about the probability distribution of the underlying random variables in each time series.
3. The algorithm is parameter-free, requiring no statistical estimation or numerical optimization.
4. The approach has low computational complexity, allowing for fast generation of scores, giving SMEs the necessary time to intervene and address any detected anomalies.
5. Unlike most ML anomaly detection models, the threshold scores and values used to classify observations as anomalies are automatically fixed, minimizing the risk of confusion and ensuring optimal algorithm performance. This approach also allows for proposing optimal threshold values for each KPI, above which the KPI becomes alarming (further clarification is provided in the application section). This is the first anomaly score system to generate scores and assist with setting optimal scoring thresholds with minimal intervention from SMEs.

Note that, to address the risk of the first records in a time series being declared as anomalies, even if their values are not high enough, a practical solution is to run the algorithm on a warm-up period before initiating the extraction and detection of anomalies.

B. Multidimensional Abnormality Score

To obtain a more comprehensive understanding of abnormal behaviors, it is preferable to develop a scoring system that takes into account all available features at a given point in time and generates an abnormality score reflecting the interaction between all variables (i.e., KPIs). Suppose that we have $l$ variables characterizing the status of a system over time, denoted by $\left\{X_t^i, t \geq 1 \text{ and i} = 1, \dots, \text{l}\right\}$. As a first step, we define upper records in a multidimensional context using the following two definitions:

1. $\forall t \geq 1$, an observation $X_t = \left(X_t^1, \dots, X_t^l\right)$ is considered to be an upper record if it is a record on at least one of the underlying dimensions. In other words, if there exists an

$\exists i \in \{1, \dots, l\}$ such that $X_t^i > \max_{j<t} X_j^i$. This definition is referred to as the "At Least One-Based Multidimensional Record" (ALO) in the rest of this paper. It is worth noting that this definition of records in a multidimensional context is introduced in Arnold et al. (2011; page 266).

2. $\forall t \geq 1$, the first step is to compute the Euclidian distance from the origin to the observation $X_t = \left(X_t^1, \dots, X_t^l\right)$:

$$d_t = \sqrt{\sum_{i=1}^{l} \left(X_t^i\right)^2}$$

Then, based on the time series $\{d_t, t \geq 1\}$ instead of $\{X_t, t \geq 1\}$, the abnormality score at time $t$ is computed in the same manner as in Subsection V-A. This approach will be called the "Distance-Based Multidimensional Record". However, this approach has a weakness in that it transforms the multidimensional data into one distance series, losing information about the impact of each underlying variable on the final abnormality score. Consequently, this approach cannot interpret the scores on a variable (KPI) level or determine the root cause of the anomaly. Since SMEs prefer models that can be used for both anomaly scoring and root cause analysis, the ALO approach will be the sole focus of the paper going forward.

Once the record series of the underlying multidimensional time series dataset has been collected using the ALO approach, the next step is to modify the abnormality score formula proposed in Subsection V-A to suit the multinational context. Let $\{X_t, t \geq 1\}$ be the multidimensional time series that displays the behavior of $l$ KPIs over time. $\forall t \geq 1$, we denote:

$$\Lambda_t = \left\{R_n = (R_n^1, \dots, R_n^l), n \geq 1 \text{ such that } R_n \text{ is the } n^{th} \text{ record of the series } \{X_i, 1 \leq i < t\}\right\},$$

$$\Lambda_t^* = \left\{R_n \in \Lambda_t \text{ such that } \exists j \in \{1, \dots, l\} \text{ with } R_n^j \geq X_t^j\right\},$$

$$\Lambda_{t,j}^* = \left\{R_n \in \Lambda_t^* \text{ such that } R_n^j \geq X_t^j\right\} \text{with } j \in \{1, \dots, l\},$$

$$\mathcal{D}_{t,j} = \left\{R_n^j - X_t^j, \text{such that } R_n \in \Lambda_{t,j}^*\right\} \text{with } j \in \{1, \dots, l\},$$

$$\overline{\mathcal{D}}_{t,j} = \begin{cases} 0, & \text{if } \mathcal{D}_{t,j} = \emptyset \\ \text{Arithmetic average of the elements of } \mathcal{D}_{t,j}, & \text{Otherwise} \end{cases}, \quad \text{with } j \in \{1, \dots, l\}.$$

Then, for each observation $X_t, t \geq 1$ the corresponding abnormality score is given by:

$$\mathcal{S}_t = \begin{cases} 1, & \text{if } X_t \text{ is a record} \\ \dfrac{1}{\left(1 + \dfrac{\text{Card } \Lambda_t^*}{\text{Card } \Lambda_t}\right)} \left(\dfrac{1}{1 + \sum_{j=1}^{l} \overline{\mathcal{D}}_{t,j}}\right), & \text{Otherwise} \end{cases}.$$

## VI. Real-World Data Application

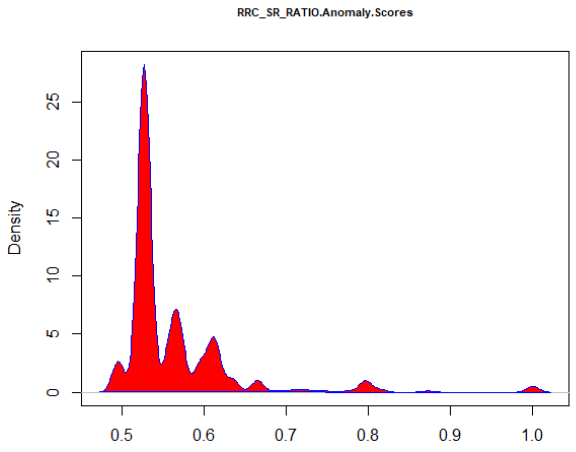
The data analyzed in this research consists of 18 primary metrics (KPIs) that assess the quality of service provided by a virtual telecommunications network cell. These KPIs are consolidated hourly, resulting in 955 observations in total, where each observation represents 1 hour of data. The KPIs include metrics such as Downlink and Uplink volume of data, Downlink and Uplink throughput, network availability, call setup success rate, and dropped call rate. These metrics

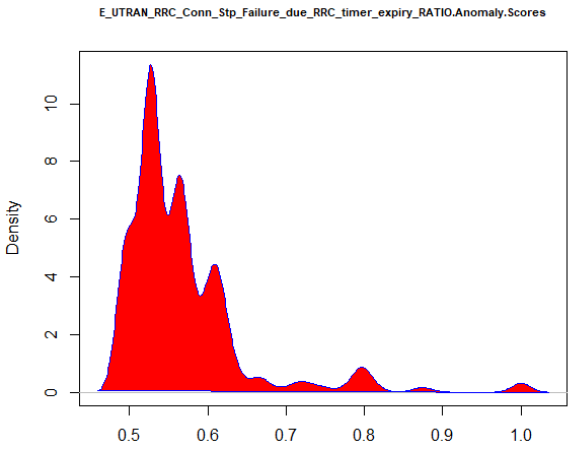
play a vital role in measuring the efficiency and effectiveness of data transmission over the network, as well as the overall performance of the cell.
Analyzing the dataset provides valuable insights into the virtual telecommunication network cell's performance and helps identify areas for improvement. For example, a high dropped call rate could indicate network congestion or other issues that need to be addressed by implementing corrective measures to enhance the quality of service offered to customers. In summary, the dataset used in this study presents a comprehensive view of the virtual telecommunication network cell's performance, empowering network operators to make informed decisions about resource allocation and optimize network performance to enhance the user experience.
To account for the specificities of telecom time series data, the KPI values have been standardized to fit within the interval [0,1] and transformed to give higher values a more alarming indication of abnormal behavior. Therefore, we are working within a space of dimensions $[0,1]^{955\times18}$, with a focus on the upper records for each of the underlying variables. It should be noted that when a KPI reaches the upper bound (e.g., KPI = 1), this observation is regarded as a new record.
Before starting anomaly scoring, a feature selection process is undertaken using the methodology described in Section IV, where only the high and medium risk KPIs are considered, following the Yang and DTRW record models, respectively. Table I shows the selected KPIs and their corresponding risk levels in terms of anomaly detection.

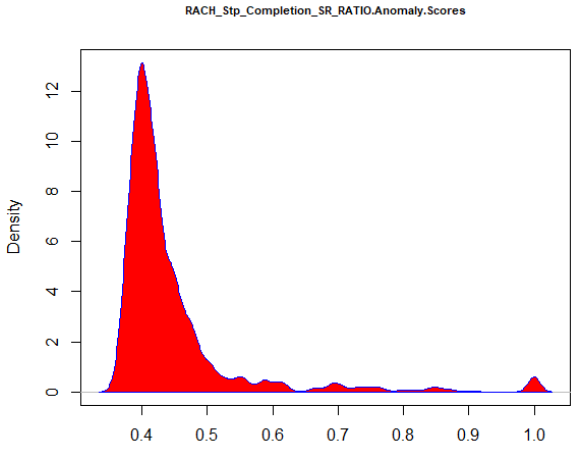
| KPI | Risk Level |
|---|---|
| RRC_SR_RATIO | High |
| E_UTRAN_RRC_Conn_Stp_Failure_due_RRC_timer_expiry_RATIO | High |
| RACH_Stp_Completion_SR_RATIO | Medium |
| Total_E_UTRAN_RRC_Conn_Stp_SR_RATIO | High |
| E_RAB_QCI1_DR_RATIO | Medium |
| DCR_LTE_RATIO | Medium |
| LTE_INTER_ENODEB_HOSR_RATIO | Medium |
| E_UTRAN_tot_HO_SR_inter_eNB_X2_RATIO | High |
| DL_THROUGHPUT_RATIO | High |
| E_RAB_DR_RATIO | Medium |

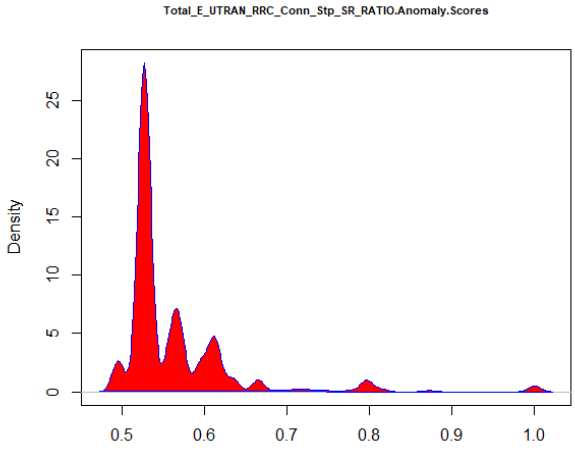
Table I: Risk level of the selected KPIs

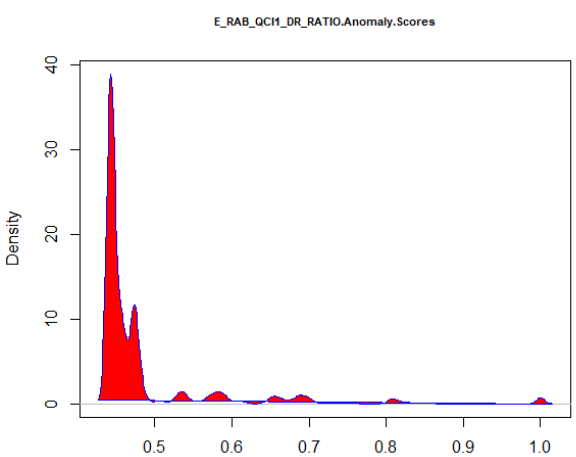
For each of the chosen KPIs, the one-dimensional abnormality score, developed in Subsection V-A, is calculated and the kernel density function of the scores is plotted in Figure II. It is evident that the probability density functions are multimodal, and that the abnormality scores associated with each of the KPIs can effectively discriminate between observations classified as normal and abnormal, using a threshold that can be determined by a simple descriptive analysis of the various distributions. Therefore, by establishing these score thresholds, an optimal corresponding KPI threshold can be recommended to SMEs to minimize classification errors. For example, consider LTE_INTER_ENODEB_HOSR_RATIO. Based on Figure II, the recommended abnormality score threshold is 0.7 (i.e., an observation with a score above 0.7 is deemed anomalous). A grid search technique is then applied to determine the optimal KPI value threshold, which is found to be 0.1131 (i.e., an observation with a KPI value above 0.1131 is considered anomalous), with a classification error rate of 5.23%. This is the first time that an anomaly detection algorithm has been able to propose a threshold for SMEs to consider, rather than the opposite. Results for all KPIs are presented in Table II.

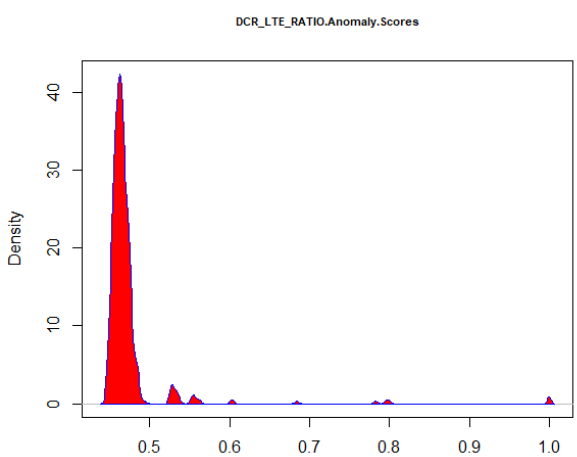

RRC_SR_RATIO.Anomaly.Scores
Density
0
5
10
15
20
25
0.5
0.6
0.7
0.8
0.9
1.0


E_UTRAN_RRC_Conn_Stp_Failure_due_RRC_timer_expiry_RATIO.Anomaly.Scores
Density
0
2
4
6
8
10
0.5
0.6
0.7
0.8
0.9
1.0


RACH_Stp_Completion_SR_RATIO.Anomaly.Scores
Density
0
2
4
6
8
10
12
0.4
0.5
0.6
0.7
0.8
0.9
1.0


Total_E_UTRAN_RRC_Conn_Stp_SR_RATIO.Anomaly.Scores
Density
0
5
10
15
20
25
0.5
0.6
0.7
0.8
0.9
1.0


E_RAB_QCI1_DR_RATIO.Anomaly.Scores
Density
0
10
20
30
40
0.5
0.6
0.7
0.8
0.9
1.0


DCR_LTE_RATIO.Anomaly.Scores
Density
0
10
20
30
40
0.5
0.6
0.7
0.8
0.9
1.0

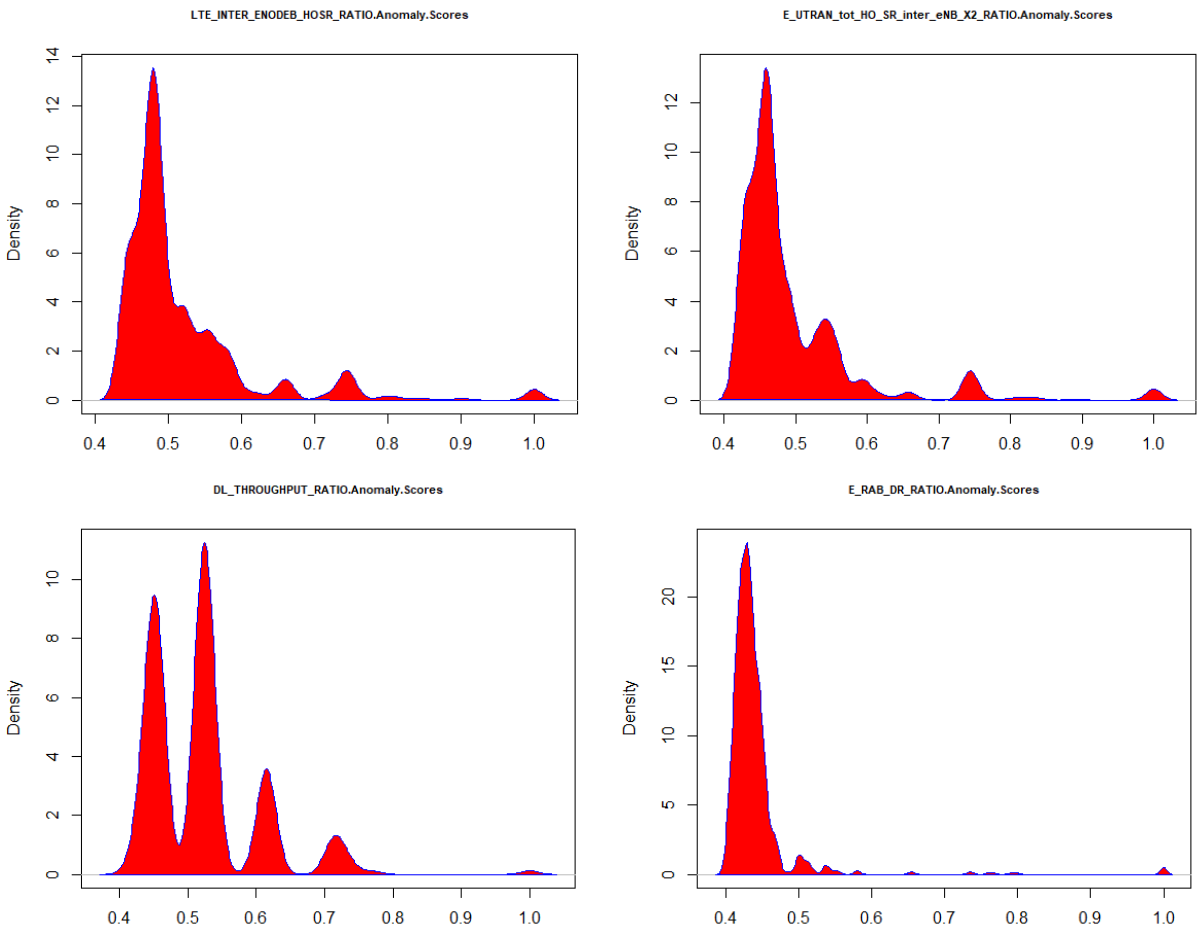



Figure II: Kernel density functions of the one-dimensional abnormality scores of each KPI.

| KPI | Score Threshold | KPI Value Threshold | Error % |
|---|---|---|---|
| RRC_SR_RATIO | 0.69 | 0.0103 | 1.47% |
| E_UTRAN_RRC_Conn_Stp_Failure_due_RRC_timer_expiry_RATIO | 0.85 | 0.0163 | 1.26% |
| RACH_Stp_Completion_SR_RATIO | 0.65 | 0.2693 | 2.83% |
| Total_E_UTRAN_RRC_Conn_Stp_SR_RATIO | 0.69 | 0.0103 | 1.47% |
| E_RAB_QCI1_DR_RATIO | 0.71 | 0.1492 | 2.30% |
| DCR_LTE_RATIO | 0.51 | 0.0854 | 0% |
| LTE_INTER_ENODEB_HOSR_RATIO | 0.7 | 0.1131 | 5.23% |
| E_UTRAN_tot_HO_SR_inter_eNB_X2_RATIO | 0 | 0.4835 | 5.13% |
| DL_THROUGHPUT_RATIO | 0.8 | 0.4414 | 0.31% |
| E_RAB_DR_RATIO | 0.6 | 0.241 | 0.21% |

Table II: One-dimensional anomaly score analysis

To assess the relationship between all the KPIs and generate a single anomaly score that represents the behavior of all the underlying variables, we will use the ALO method discussed in Subsection V-B. The graph in Figure III shows the kernel density function of the abnormality scores that were calculated. This distribution is bimodal, making it easy to distinguish between anomalies and non-anomalies without the need to estimate a threshold, unlike traditional anomaly detection models. Using this method, we identified 4.4% of observations as abnormal.

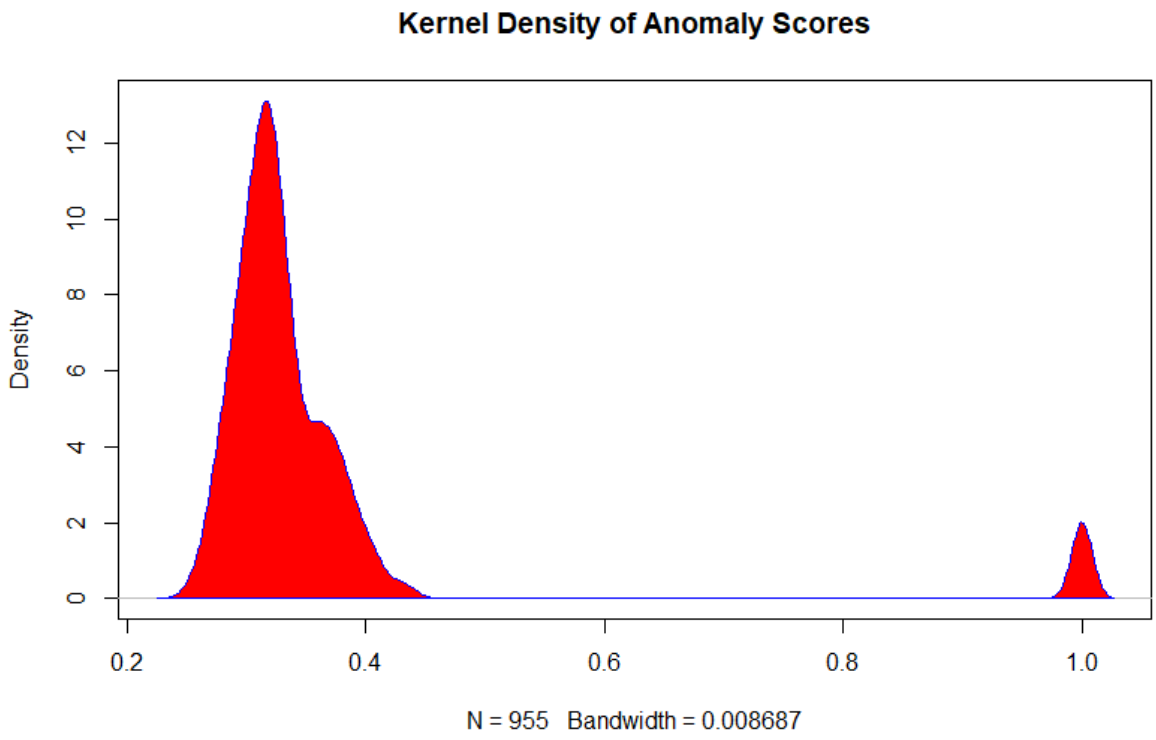



Figure III: Kernel density functions of the ALO approach abnormality scores of each KPI.

## VII. Conclusion

This paper describes the use of records theory to create two methods. The first method reduces the number of variables in a time series to focus on those that have a significant impact on abnormal behavior. The second method proposes a scoring system for anomaly detection that can be applied in one or multiple dimensions. This system can objectively detect anomalies and suggest threshold values for KPIs without the need for expert input.

The suggested anomaly detection scoring system is a simple algorithm that does not rely on any specific distribution or parameters. It is designed to be used as an online system for detecting anomalies with minimal computational complexity, and it eliminates the risk of overfitting. Additionally, the system can automatically estimate the threshold value needed to classify observations as anomalies, ensuring optimal performance of the algorithm. Furthermore, the algorithm was tested on real-world telecommunications data, and it demonstrated excellent performance in detecting anomalies with very low error rates.

One possible application of this work is to conduct a more in-depth analysis of the anomaly scores in order to extract information about the underlying causes of the anomalies. Another potential direction is to explore the probabilistic properties of the different anomaly scores generated by the system, using records theory as a basis for analysis. This approach could be informed by the research conducted by Hoayek and Ducharme in 2017.

## References

[1] AKOGLU, L., TONG, H., AND KOUTRA, D. Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29 (2015), 626–688.

[2] ALI, W. A., MANASA, K., BENDECHACHE, M., FADHEL ALJUNAID, M., AND SANDHYA, P. A review of current machine learning approaches for anomaly detection in

network traffic. Journal of Telecommunications and the Digital Economy 8, 4 (2020), 64–95.
[3] ALVI, A. M., SIULY, S., AND WANG, H. Developing a deep learning-based approach for anomalies detection from EEG data. In Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I (2022), Springer, pp. 591–602.
[4] ARNOLD, B. C., BALAKRISHNAN, N., AND NAGARAJA, H. N. Records. John Wiley & Sons, 2011.
[5] BALLERINI, R., AND RESNICK, S. Records from improving populations. Journal of Applied Probability 22, 3 (1985), 487–502.
[6] CHANDLER, K. The distribution and frequency of record values. Journal of the Royal Statistical Society: Series B (Methodological) 14, 2 (1952), 220–228.
[7] CHANDOLA, V., BANERJEE, A., AND KUMAR, V. Anomaly detection: A survey. ACM computing surveys (CSUR) 41, 3 (2009), 1–58.
[8] HAMIE, H., HOAYEK, A., AND AUER, H. Modeling the price dynamics of three different gas markets-records theory. Energy Strategy Reviews 21 (2018), 121–129.
[9] HOAYEK, A. S., DUCHARME, G. R., AND KHRAIBANI, Z. Distribution-free inference in record series. Extremes 20, 3 (2017), 585– 603.
[10] HUANG, Z., KANG, X., LI, S., AND HAO, Q. Game theory-based hyperspectral anomaly detection. IEEE Transactions on Geoscience and Remote Sensing 58, 4 (2019), 2965–2976.
[11] JABBOUR, C., HOAYEK, A., MAUREL, P., KHRAIBANI, Z., AND GHALAYINI, L. Examining satellite images market stability using the records theory: Evidence from French spatial data infrastructures. Journal of Spatial Information Science, 22 (2021), 61–82.
[12] KHRAIBANI, Z., JACOB, C., DUCROT, C., CHARRAS-GARRIDO, M., AND SALA, C. A non-parametric exact test based on the number of records for an early detection of emerging events: illustration in epidemiology. Communications in Statistics-Theory and Methods 44, 4 (2015), 726–749.
[13] LJUNG, G. M., AND BOX, G. E. On a measure of lack of fit in time series models. Biometrika 65, 2 (1978), 297–303.
[14] MAJUMDAR, S. N., AND ZIFF, R. M. Universal record statistics of random walks and lévy flights. Physical Review Letters 101, 5 (2008), 050601.
[15] NEVZOROV, V. B. Records. Theory of Probability & Its Applications 32, 2 (1988), 201–228.
[16] NEVZOROV, V. B. Records: mathematical theory. American Mathematical Soc., 2001.
[17] RASHID, A. B., AHMED, M., SIKOS, L. F., AND HASKELLDOWLAND, P. Anomaly detection in cybersecurity datasets via cooperative co-evolution-based feature selection. ACM Transactions on Management Information Systems (TMIS) 13, 3 (2022), 1–39.
[18] ŠABIC´, E., KEELEY, D., HENDERSON, B., AND NANNEMANN, S. Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data. AI & SOCIETY 36, 1 (2021), 149–158.
[19] SALTON, G., WONG, A., AND YANG, C.-S. A vector space model for automatic indexing. Communications of the ACM 18, 11 (1975), 613–620.
[20] SHA, W., ZHU, Y., CHEN, M., AND HUANG, T. Statistical learning for anomaly detection in cloud server systems: A multi-order Markov chain framework. IEEE Transactions on Cloud Computing 6, 2 (2015), 401–413.
[21] VANGIPURAM, R., GUNUPUDI, R. K., PULIGADDA, V. K., AND VINJAMURI, J. A machine learning approach for imputation and anomaly detection in IoT environment. Expert Systems 37, 5 (2020), e12556.
[22] WERGEN, G. Records in stochastic processes theory and applications. Journal of Physics A: Mathematical and Theoretical 46, 22 (2013), 223001.

[23] WERGEN, G., AND KRUG, J. Record-breaking temperatures reveal a warming climate. EPL (Europhysics Letters) 92, 3 (2010), 30008.
[24] ZHOU, X., HU, Y., LIANG, W., MA, J., AND JIN, Q. Variational LSTM enhanced anomaly detection for industrial big data. IEE