Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-026-00667-8

REGULAR ARTICLE

Fast, robust, and accurate anomaly detection for multivariate time series

Simone Tonini[1] · **Andrea Vandin**[1,2] · **Francesca Chiaromonte**[1,3] · **Daniele Licari**[1] · **Fernando Barsacchi**[4]

Received: 20 May 2025 / Revised: 24 December 2025 / Accepted: 3 January 2026
© The Author(s) 2026

Abstract

Anomaly detection is a topic widely studied both in Statistics and Computer Science, with an ever growing literature in both disciplines. We present a novel, fast, robust, accurate, and widely applicable semi-supervised procedure for anomaly detection in multivariate time series, FRA^2Nk (Fast, Robust, and Accurate ANomaly detection). It comprises 5 steps: smoothing, multicollinearity mitigation, dissimilarity measurement, threshold selection, identification of the causes of the anomalies. FRA^2Nk can tackle issues from different challenging contexts, where signals can be highly multicollinear, have unknown distributions, and intertwine short-lived noise with longer-lived anomalies. Using several experiments, we demonstrate the generality, low computational cost, precision, and interpretability of FRA^2Nk. In particular: (i) Using public benchmark datasets from anomaly detection, we evaluate the computational cost and performance of FRA^2Nk against the semi-supervised methods from a recent literature review, finding that FRA^2Nk is effective, broadly applicable, and that it outperforms existing approaches in anomaly detection and runtime; (ii) Using such datasets we also show that FRA^2Nk can explain the causes of the discovered anomalies; (iii) Using simulation studies, we show that FRA^2Nk is robust to several possible issues in the data; (iv) Using a case study from an industrial partner, we show that FRA^2Nk is effective.

Keywords Anomaly detection · Multivariate time series · Semi-supervised methods

Simone Tonini, Andrea Vandin, Francesca Chiaromonte, Daniele Licari, and Fernando Barsacchi contributed equally to this work.

Extended author information available on the last page of the article

Published online: 30 January 2026

1 Introduction

Nowadays, the use of data is crucial in many domains. For example, manufacturers collect massive amounts of data on many aspects of their production processes – usually in the form of time series concerning a large number of variables. This data helps domain experts detect potentially *anomalous* behaviors (production errors, technical problems, system defects, breakdowns, outages) which can make processes inefficient or sub-optimal. A rich literature on anomaly detection exists, comprising contributions from both Statistics and Computer Science (see, e.g., Schmidl et al. (2022); Choi et al. (2021) for reviews). In particular, here we use the exhaustive review by Schmidl et al. (2022) as a reference to frame our proposal in the context of this vast literature. Notwithstanding the plethora of contributions to date, anomaly detection remains a particularly complex task (Schmidl et al. 2022). The main difficulty lies in identifying approaches that guarantee at the same time broad applicability, good performance, and low computational cost. Several characteristics of time series can create challenges in anomaly detection. For example, we may have confounding environmental effects (e.g., vibrations, temperature, humidity in industrial settings) which induce a variability of the same magnitude as that due to actual anomalies, complicating identification of the latter (see, e.g., Deraemaeker and Worden (2018); Grosskopf et al. (2022)). This may limit the range of anomaly detection methods that can be successfully implemented in practice.

Our contribution consists of a novel semi-supervised anomaly detection procedure, named FRA^2Nk (Fast, Robust, and Accurate ANomaly detection), which leverages a combination of relatively simple statistical tools (smoothing filters, variance inflation factors, the Mahalanobis distance, threshold selection algorithms from extreme value theory, and feature importance techniques). We evaluate performance and computational cost by comparing our procedure with state-of-the-art semi-supervised methods reviewed in Schmidl et al. (2022). The comparison is conducted on 8 widely used public benchmark datasets from the anomaly detection literature, representing a broad range of domains and heterogeneous applications. We employ common performance metrics such as the Matthews correlation coefficient (MCC) (Chicco and Jurman 2020), F1, recall, and precision. We also assess computational cost through runtime analyses. Overall, we conclude that our procedure is effective, broadly applicable, and outperforms existing approaches in both performance in anomaly detection and runtime. In addition to benchmark datasets we also consider a concrete case study offered by an industrial partner, showing the effectiveness in practice of FRA^2Nk. It consists of a multivariate time series containing 70,000 observations for 119 variables collected on a production line from the paper and nonwoven industry.

Our goal is to develop a methodology to detect anomalies in scenarios with two key general requirements, namely: (**R1**) enable quick and accurate identification of the time intervals in which anomalies occur; (**R2**) provide information on patterns and variables involved in the anomalies, to help domain experts decipher their causes. In order to lower the burden on users in labeling data, we assume that the data at our disposal contains exclusively intervals assessed as *anomaly-free* by the domain experts. In other words, we cannot resort to supervised methods, which need to be trained on data comprising both anomalies and anomaly-free observations. The

procedure we developed thus classifies as *semi-supervised* based on the definition by Schmidl et al. (2022); we do not exploit *anomaly* labels in the training data, but rather data concerning only *normal*, anomaly-free behavior.

In addition to **R1** and **R2**, we also offer support for domains where measurements may be affected by short-term noise (e.g., errors in sensor measurements due to dust or environmental conditions). For this, FRA^2Nk has an optional initial step to allow users to focus on *long-lived anomalies* (e.g., anomalies lasting for minutes over observations taken per second).[1]

The last decade has seen the development of a multitude of sophisticated semi-supervised methods based, e.g., on deep learning or statistical learning methods, which should be able to tackle **R1-2** by leveraging complex patterns in the data without the introduction of strong assumptions (see Choi et al. (2021); Blázquez-García et al. (2021); Schmidl et al. (2022) for recent reviews). However, the comprehensive comparison in Schmidl et al. (2022) demonstrates that no single method offers the best performance across different scenarios. In contrast, FRA^2Nk exploits the *Mahalanobis distance* (Mahalanobis 1936) evaluated on the training data as a means to identify anomalies. This requires only estimates of location and covariation from the training data. However, in order to obtain a robust and general procedure, it is necessary to address technical challenges (**TC**) that can often be found in multivariate time series. The first (**TC1**) is multicollinearity. Figure 1 (a) shows the eigenvalues of the correlation matrix of the 119 variables of our case study; the smallest are close to 0, indicating strong multicollinearity. This is a problem, since the Mahalanobis distance requires the covariance matrix to be invertible. The second technical challenge (**TC2**) concerns the distribution of the data. Figure 1 (b) shows the densities of 5 variables randomly selected among the 119. Clearly, even under anomaly-free conditions, the distributions are far from Gaussian, or other typical forms, e.g., used to represent skewed data (see, e.g., Hubert and Van der Veeken (2008); Tiku et al. (2010)). Nota-

[1] Note that we use the terms short-lived and long-lived anomalies to denote, respectively, anomalous

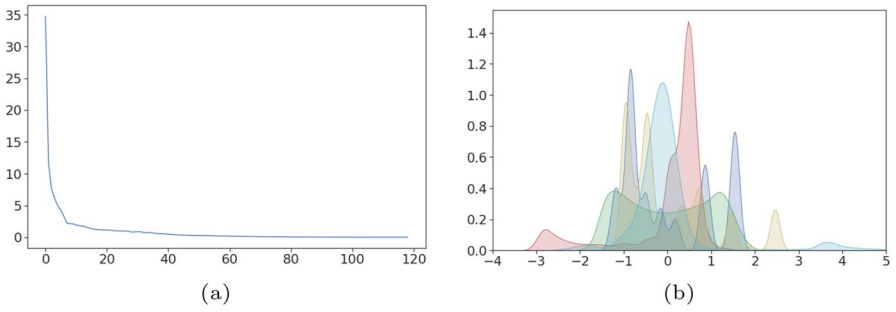

Fig. 1 Panel (**a**) shows the eigenvalues of the correlation matrix of the 119 variables considered in our case study. Panel (**b**) shows the densities of 5 among these variables after standardization

events that persist for a single or few observations and those that extend over a longer time span. These notions align with the concepts of additive outliers and temporary outliers commonly used in the time series literature (see, e.g., Bellini (2016); Kim et al. (2022)).

bly, in Supplementary Material S1 we show that the observed non-Gaussianity in the data is not just due to temporal dependence. When anomaly-free data follow a Gaussian or other known distribution, it is possible to derive a distribution for the Mahalanobis distance and to define a threshold for the identification of anomalies (see, e.g., Maronna et al. (2019). When distributional assumptions are unsuitable however, one must resort to different approaches to define an appropriate threshold.

Our procedure, which addresses all the challenges and considerations raised thus far, consists of the 5 steps depicted in Fig. 2. We start with an anomaly-free dataset, which represents the training set for our procedure. The first step consists of smoothing the data through a median or mean filter, as to remove short-lived anomalies; this step is optional. The second step mitigates multicollinearity removing variables based on their *variance inflation factors* (Craney and Surles 2002). In the third step, we compute the *Mahalanobis distances* for all points in the training set. The fourth step provides a *threshold* for the identification of anomalies. If no distributional assumptions are suitable, we use the maximum value of the Mahalanobis distances in the training set (MVT), or the value produced by a Peaks-Over-Threshold (POT) analysis of the right tail of such distances (see, e.g., (Balkema and De Haan 1974; Pickands III 1975)). The data in the test set undergoes smoothing, as the training data. A Mahlanobis distance is produced for each test data point, and a (binary) anomaly prediction is produced based on whether its distance is above threshold. Finally, the fifth step uses feature importance techniques offered by random forest (Breiman 2001) or logistic regression (James et al. 2013, ch. 4) to analyze anomalies predicted in the test set, unveiling patterns and pinpointing relevant variables, to decipher their putative causes.

Our work complements the vast literature on anomaly detection in multivariate time series. Fibbi et al. (2024) and Cerioli and Perrotta (2014) addressed the problem of robust clustering for large datasets in the presence of outliers. Vinue and Epifanio (2021) proposed the use of robust functional archetypes and adjusted boxplots as a methodology for identifying functional outliers. Zeller et al. (2019) extended finite mixture regression models for censored data using mixtures of normal distribution scales. Talagala et al. (2021) proposed an unsupervised algorithm for detecting anomalies in high-dimensional data. Raymaekers and Rousseeuw (2021) provided robust estimators of covariance matrices for ultra-high-dimensional data.

Furthermore, several recent proposals for anomaly detection focus on large or streaming data, requiring real-time data analysis. For example, Fisch et al. (2022) proposed a methodology for real-time anomaly detection designed for univariate

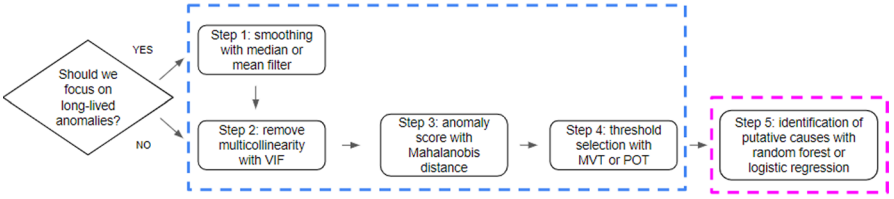

Fig. 2 Flow chart of the proposed FRA^2Nk procedure. The dashed blue line shows the steps used to tackle **R1**, while the dashed purple line shows the step used to tackle **R2**

time series. Our work is aligned with these efforts, in that we offer a computationally lean, fast procedure that can work for very large time series – and in fact is applicable to both univariate and multivariate time series.

A preliminary and limited version of FRA^2Nk was introduced in Tonini et al. (2023), where the methodology was only sketched, results were reported on a single dataset, and no comparison with existing approaches was provided.

The remainder of the article is organized as follows. Section 2 focuses on the properties of the Mahalanobis distance. Section 3 details the FRA^2Nk procedure. Section 4 demonstrates its performance through a comprehensive validation on methods and datasets from Schmidl et al. (2022). Section 5 relates its feature importance results to the causes of the anomalies. In Sect. 6 we validate the impact of temporal dependence and of anomalies in the training sample by means of simulated data. Finally, Sect. 7 demonstrates FRA^2Nk on the case study, and Sect. 8 provides final remarks. Code and replicability material for Sects. 4 and 5 is available at https://zenodo.org/records/15076198.

2 Some background on the mahalanobis distance

The Mahalanobis distance (Mahalanobis 1936) is one of the key statistical tools employed by FRA^2Nk. It has low computational costs and high interpretability, as it allows one to quantify how much an observation deviates from "behavior under normal conditions" through straightforward estimation of the mean vector and covariance matrix of the training data. Normal conditions refer to the state in which a system or process operates within its expected, designed, or acceptable parameters, without faults, anomalies, or disturbances. It represents the baseline or reference conditions against which any deviation can be interpreted as a potential issue (e.g. fault, degradation, or abnormal event). These conditions are usually defined by specific ranges for factors such as temperature, pressure, humidity, and sometimes even air quality and ventilation (see Deraemaeker and Worden (2018); Huang et al. (2022)). In complex multivariate systems such as industrial processes, normal conditions may encompass substantial environmental and operational variability. Within this context, an anomaly is defined as one or more observations that fall outside the expected variability of these conditions. In particular, we specifically focus on point anomalies (see Choi et al. (2021)), a data point or sequence that deviates sharply from the normal behavior.

Here, we focus on aspects that make the Mahalanobis distance suitable for the detection of point anomalies in semi-supervised contexts, while we discuss other background material in Supplementary Material S5. Let $\mathrm{X}_A = \{\mathrm{x}_t\}_{t=1}^{T_A}$ denote an $n \times T_A$ rectangular array of observations on n weakly stationary time series with T_A observations for each series. In practice, X_A is provided by a domain expert to denote the behavior of the process under normal conditions. Given a new rectangular array $\mathrm{X}_B = \{\mathrm{x}_t\}_{t=T_A+1}^{T}$ of dimension $n \times T_B$, where $T_B = T - T_A$, we focus on the question: *How distant is each point in* X_B *from the behavior captured by* X_A? The Mahalanobis distance is a common approach to answer this. Let $\boldsymbol{\mu}_A$

and $\boldsymbol{\Sigma}_A$ be the population mean vector and covariance matrix of X_A, and let $\widehat{\boldsymbol{\mu}}_A$ an $\widehat{\boldsymbol{\Sigma}}_A$ be their estimates. We have that $\widehat{\boldsymbol{\mu}}_A \xrightarrow{a.s.} \boldsymbol{\mu}_A$ and $\widehat{\boldsymbol{\Sigma}}_A \xrightarrow{a.s.} \boldsymbol{\Sigma}_A$ at convergence rate $O_p\left(\sqrt{\frac{1}{T_A} \cdot \left\|\sum_{k=-\infty}^{\infty} \Gamma(k)\right\|}\right)$, where $\Gamma(k) = \mathrm{Cov}(\mathrm{X}_t, \mathrm{X}_{t+k})$ indicates the autocovariance matrices.[2] For a given n-dimensional observation in the test set, say $\mathrm{x}_{\tilde{t}} = (x_{1,\tilde{t}}, \ldots, x_{n,\tilde{t}})'$ with $T_A+1 \leq \tilde{t} \leq T$, the Mahalanobis distance from the training set is

$$MD_{\tilde{t}} = \sqrt{(\mathrm{x}_{\tilde{t}} - \widehat{\boldsymbol{\mu}}_A)' \, \widehat{\boldsymbol{\Sigma}}_A^{-1} \, (\mathrm{x}_{\tilde{t}} - \widehat{\boldsymbol{\mu}}_A)} \quad . \tag{1}$$

$MD_{\tilde{t}}$ measures how far $\mathrm{x}_{\tilde{t}}$ is from the center of the training data in an inner product that is shaped by its inverse covariance – so directions of higher (co)variation matter less in assessing departure from the center. As mentioned in the Introduction, industrial processes can have large operational variability, as well as collinearities, under normal conditions. The next assumption is fundamental to apply semi-supervised anomaly detection methods.

Assumption 1 X_A is representative of the operational variability of the process under normal conditions.

Remark 1 Under Assumption 1, X_A should comprehensively capture the range of normal operating regimes that the system may experience. To this end, T_A must be sufficiently large to encompass recurring temporal patterns, such as seasonal or cyclic behaviors, that characterize normal operation.

In the context of point anomalies, Assumption 1 is empirically expected to hold when the extreme values in the training data – reflecting normal operational variability – are substantially lower than the anomalous values in the test data for the same variables.

Assumption 1 is crucial for preventing the masking effect, which occurs when true anomalies remain undetected because they are blended by other extreme values or by the inherent variability of the normal data. Also, Assumption 1 supports the effective application of semi-supervised methods for detecting point anomalies in X_B, reducing the risk of swamping, which occurs when normal observations are erroneously classified as anomalies.[3]

Under Assumption 1 we can identify $p \leq n$ eigenvectors of $\widehat{\boldsymbol{\Sigma}}_A$ whose span approximates the linear sub-space to which the normal conditions belong. In particular, let $\mathrm{v}_1, \ldots, \mathrm{v}_n$ be the eigenvectors of $\widehat{\boldsymbol{\Sigma}}_A$ ordered based on the corresponding

[2] Under weak dependence, fast decay of autocovariances leads to $O_P(T_A^{-1/2})$, while slow decay (e.g., long-range dependence) results in slower convergence. (see, e.g., Hamilton 1994, ch. 3).

[3] Note that, in the context of point anomalies in semi-supervised settings, masking and swamping tend to arise together only when both of the following conditions are met: (*i*) the training set contains anomalies with absolute values greater than those in the test set (which leads to masking), and (*ii*) the test set contains extreme values that exceed those in the training set and are falsely flagged as anomalies (which leads to swamping).

eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, and let p be the smallest integer such that $\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \lambda_i} > \alpha$, where $\alpha \in [0, 1]$ is a threshold value (e.g., 0.99). This means that the first p principal components explain $100 \times \alpha\%$ of the variability of X_A. Under Assumption 1, such components explain most of the operational variability of the process under normal conditions, while the Mahalanobis distance will increase more steeply for an observation varying in the direction of the remaining $n - p$ components (Deraemaeker and Worden 2018). The following proposition formalizes this reasoning.

Proposition 1 Under Assumption 1, let $\sum_{i=1}^{p} \lambda_i$ be the sum of the p largest eigenvalues, associated with the eigenvectors that explain $100 \times \alpha\%$ of the variability of X_A. If $\sum_{i=1}^{p} \lambda_i \to \sum_{i=1}^{n} \lambda_i$ (i.e., $\alpha \to 1^-$), then $MD_{\tilde{t}}$ tends to the Mahalanobis distance of $\mathrm{x}_{\tilde{t}}$ projected on the space described by the remaining $n - p$ components.

See Supplementary Material S2 for the proof. Proposition 1 shows that by including all operational variability (the variables measured in all possible environmental conditions) in computing $\widehat{\boldsymbol{\Sigma}}_A$, the Mahalanobis distance increases as $\mathrm{x}_{\tilde{t}}$ moves away from the normal conditions of the process. This property justifies its use in anomaly detection. However, as discussed in the Introduction, some technical challenges (**TC1-2**) may complicate such use in the considered domain. Section 3 shows how some of the step included in FRA^2Nk solve these issues.

3 The 5-step FRA^2Nk procedure

In this section, we present our procedure to detect anomalies in industrial contexts. FRA^2Nk extends the traditional application of Mahalanobis distance to the broader and more complex setting of multivariate time series, relying solely on Assumption 1. This extension introduces key challenges, particularly the unknown data distribution and high multicollinearity among variables. Although numerous deep learning and machine learning techniques can address these challenges, we adopt a selection of well-established techniques. Our selection is guided by simplicity: we select well-known techniques that empirical researchers and practitioners can easily implement and that can be generalized to many contexts. In particular, FRA^2Nk combines five main elements; namely: smoothing filters, variance inflation factors, Mahalanobis distance, threshold selection procedures, and feature importance techniques. Here, we present how these components are employed throughout the various steps of FRA^2Nk. Where appropriate, we also illustrate how the methodology can be extended and adapted to address specific problems.

Let X_A and X_B be rectangular arrays on n variables as in Sect. 2. For the $n \times T$ matrix $\mathrm{X} = (\mathrm{X}_A, \mathrm{X}_B)$, the proposed procedure comprises the following steps.

Step 1 - Smoothing (optional)

If the process analyst is interested only in long-lived anomalies, we smooth each time series with a filter based on location measures (mean, or median for robustness) computed with a moving window of size h. This replaces each of the T-dimensional vectors x_i, $i = 1, \ldots, n$ with a corresponding $(T - (h - 1))$-vector of smoothed val-

ues w_i. To simplify notation, we indicate with $\ddot{T} = T - 2(h-1), \ddot{T}_A = T_A - (h-1)$ and $\ddot{T}_B = T_B - (h-1)$ the sizes of the smoothed vectors. Therefore, we replace the $n \times T$ matrix $\mathrm{X} = (\mathrm{X}_A, \mathrm{X}_B)$ with the $n \times \ddot{T}$ matrix $\mathrm{W} = (\mathrm{W}_A, \mathrm{W}_B)$, where W_A ($n \times \ddot{T}_A$) comprises smoothed values in normal conditions, and W_B ($n \times \ddot{T}_B$) smoothed values among which we want to detect anomalies.

Smoothing allows us to mitigate short-lived anomalies that would be considered as sensor noise by domain experts, and thus to avoid false positives. Domain experts will be in a position to select between mean filtering or the more robust median one, and to specify an appropriate window size *h* for given applications. In the experiments in the following Sections, we rely on median filtering and we consider either no smoothing ($h = 1$), or smoothing with $h = 10$. This choice of smoothing parameter gave good results on all datasets considered in our experiments, but in Supplementary Material S4.1 we study the performance of FRA^2Nk for different values of *h*. Also, in Supplementary Material S4.2 we show that choosing a sufficiently large value of *h* and using the median as a filter (rather than the mean) allows us to effectively handle persistent anomalies caused by high temporal dependence.

Step 2 - Removing multicollinearity (to address **TC1**)

We use Variance Inflation Factors, VIF (Craney and Surles 2002), to quantify multicollinearity among the variables in the training data W_A (or X_A if we omit the smoothing) and iteratively remove some of them. We first subtract from the entries of each w_{Ai} its mean $\bar{w}_{Ai} = \ddot{T}_A^{-1} \sum_{t=1}^{\ddot{T}_A} w_{it}$, switching to $\mathrm{d}_{Ai} = \mathrm{w}_{Ai} - \bar{w}_{Ai}1$, and form the centered matrix D_A (centering is not necessary here; however, popular statistical software packages require it in order to compute VIFs ignoring intercepts). Next, we regress each variable in D_A against all others, compute the coefficients of determination from such regressions, say R_i^2, and thus the VIFs, which are given by $\mathrm{VIF}_i = 1/(1 - R_i^2)$, $i = 1, \ldots n$. We remove the variable with largest VIF, re-run the regressions and recompute the VIFs, remove again the variable with largest VIF, *et cetera* – until the largest VIF is less than 5, a benchmark commonly used in the literature (James et al. 2013, ch. 3.3). This leaves us with a set of m ($\leq n$) variables with at most mild multicollinearity. We indicate the reduced, centered training data matrix as $\widetilde{\mathrm{D}}_A$.

Mitigating multicollinearity is critical because the Mahalanobis distance utilizes the inverse of the covariance matrix estimated on the data (see Equation (1)); the strong linear associations which exist in our case study (see Fig. 1d) and many other applications may prevent such inversion. The covariance matrix $\tilde{\boldsymbol{\Sigma}}_A = \ddot{T}_A^{-1} \widetilde{\mathrm{D}}_A \widetilde{\mathrm{D}}_A'$ relative to the *m* non-multicollinear variables will not present any invertibility issues.

Note that temporal dependence among regressors can lead to increased multicollinearity. This, in turn, inflates the variance inflation factors (VIFs), since VIF is directly related to the degree of linear dependence among predictors. As a consequence, when standard VIF thresholds (e.g., VIF < 5) are used as criteria for variable selection, fewer regressors may be retained in the presence of temporal dependence compared to the case of independence. This does not pose an issue for FRA^2Nk, provided that the selected variables remain representative of the operational variability of the process under normal conditions.

In settings where the number of variables exceeds the number of observations ($n > \ddot{T}_A$), traditional VIF calculations become inapplicable. However, FRA^2Nk can be readily adapted to high-dimensional settings by replacing VIF with appropriate dimensionality-reduction techniques, such as principal component analysis, which remain applicable when the number of variables exceeds the number of observations.

Step 3 - Computing anomaly scores

We start by processing the test data W_B. First, we center it with the means computed on the training data, switching to $\mathrm{d}_{Bi} = \mathrm{w}_{Bi} - \bar{w}_{Ai}1$ and forming the centered matrix D_B. Next, we reduce such matrix eliminating the same variables that were eliminated from the training data in Step 2; we indicate the reduced, centered test data matrix as $\widetilde{\mathrm{D}}_B$. Finally, we calculate Mahalanobis distances for the full dataset $\widetilde{\mathrm{D}} = (\widetilde{\mathrm{D}}_A, \widetilde{\mathrm{D}}_B)$ (that is, for both training and test observations) using the training covariance matrix; in symbols $MD_t = \sqrt{\widetilde{\mathrm{d}}_t' \tilde{\boldsymbol{\Sigma}}_A^{-1} \widetilde{\mathrm{d}}_t}$ $t = 1, \ldots, \ddot{T}$. We indicate with $\boldsymbol{MD}_T$, $\boldsymbol{MD}_A$ and $\boldsymbol{MD}_B$, respectively, the T, $\ddot{T}_A$ and $\ddot{T}_B$-dimensional vectors of anomaly scores for the full dataset, the training set and the test set.

This step relies on the classical (non-robust) Mahalanobis distance. While this choice might appear restrictive, Proposition 1 demonstrates that, under Assumption 1, the classical Mahalanobis distance remains a highly effective tool for anomaly detection in semi-supervised settings.

Nevertheless, FRA^2Nk is inherently flexible and can be readily extended to more complex scenarios in which the training data may exhibit non-elliptical distributions, outliers, or contamination. Assumption 1 no longer holds. In such cases, a filtering step is required to ensure the reliability of semi-supervised methods. This can be achieved through robust statistical techniques, as the popular Minimum Covariance Determinant (MCD) estimator (see, e.g., Kalina and Tichavskỳ (2022)). The MCD identifies the subset of observations whose covariance matrix has the smallest determinant, thereby enabling the computation of a robust Mahalanobis distance. This approach provides more reliable estimates of the covariance matrix $\tilde{\boldsymbol{\Sigma}}_A$, even in the presence of non-elliptical distributions, outliers, or data contamination.

For an overview of the MCD estimator, including its desirable properties such as affine equivariance, breakdown value, and influence function, see Hubert and Debruyne (2010). Moreover, recent advancements have extended the applicability of MCD to high-dimensional settings. Boudt et al. (2020) propose the Minimum Regularized Covariance Determinant (MRCD) estimator, which incorporates regularization to ensure positive definiteness of the covariance matrix, even when the number of variables significantly exceeds the sample size. This extension makes robust Mahalanobis-based filtering feasible in high-dimensional applications.

Notably, in the context of time series, the applicability of the Mahalanobis distance depends on the stability of the temporal dependence structure. Under Assumption 1, the training sample captures the natural temporal dependence patterns, thereby enabling reliable estimation of the covariance structure and thresholds. However, if the temporal dependence evolves, adaptive or dynamic modeling approaches may be necessary to maintain detection accuracy. For instance, FRA^2Nk could incorporate a step that analyzes the evolution of the autocorrelation function (ACF) over time.

Once time intervals are identified in which the ACF remains approximately stable, separate Mahalanobis distances can be computed for each interval. The most suitable distance would then be selected as the one that best reflects the ACF structure observed in the test sample.

Step 4 - Threshold selection (to address **TC2**)

We use $\boldsymbol{MD}_A$ to select a threshold k beyond which an observation in $\boldsymbol{MD}_B$ is flagged as an anomaly. Note that if we can assume the variables retained in Step 2 to be distributed as an m-variate Gaussian under normal conditions, the quadratic form expressing the Mahalanobis distance will be distributed as a chi-square with m degrees of freedom (Penny 1996; Tomarchio and Gallaugher 2024). Thus, setting the threshold as a quantile of such chi-square, say $k = \chi^2_m(1-\alpha)$, would guarantee a p-value of α when flagging anomalies (one-sided rejections). However, unfortunately, the variables in our reference domain are far from Gaussian, as well as from other forms of known distribution. (see Fig. 1 e). This prevents the direct computation of theoretical p-values for test observations. One might consider using empirical p-values as an alternative. However, empirical p-values are most informative when accompanied by an analytical null model, where procedures such as Benjamini–Hochberg or scan statistics can ensure rigorous false discovery rate (FDR) control (see, e.g., Noble (2009); Arias-Castro et al. (2018)). In contrast, when the null distribution itself is estimated empirically, p-values become a redundant intermediate step, offering little to no advantage for FDR estimation and potentially introducing additional estimation noise. This consideration motivates the adoption of alternative approaches that directly model or approximate the null distribution without relying on analytically derived p-values. We therefore resort to other approaches to select the threshold k; namely, the *Maximum Value in the Training sample* (MVT) and *Peaks-Over-Threshold* (POT) approaches. In the former, the threshold is simply set at the largest value in $\boldsymbol{MD}_A$, using, in a way, an empirical p-value of 0 when flagging anomalies.

In the latter, we fit a generalized Pareto distribution (Siffer et al. 2017) to the *peaks* in the training set, i.e., the elements of $\boldsymbol{MD}_A$ above a given percentile (we fix the 99th one)[4]. Using a known formula (see Supplementary Material S5.1), we thus obtain a threshold k that depends on the peaks, on the parameters of the fitted distribution, and on a chosen probability (we fix 0.001) for a peak to be an anomaly, or an *extreme event* in POT terminology.[5] Once the threshold k is fixed, we compare each observation $MD_{\tilde{t}}$ in $\boldsymbol{MD}_B$ with k, and generate a binary vector $\widehat{\mathrm{y}} = (\widehat{y}_1, \ldots, \widehat{y}_{\ddot{T}_B})$, with $\widehat{y}_{\tilde{t}} = 1$ if $MD_{\tilde{t}} > k$ and 0 otherwise.

Notably, the seminal article on Extreme Value Theory (EVT) ((Gnedenko 1943)) provided results for i.i.d. random variables. However, later works generalized such results to the case of dependent processes (see, e.g., Loynes (1965); Leadbetter (1974)). Based on such generalizations, POT relies on two fundamental assumptions: (i) stationarity of the underlying process, and (ii) independence of extreme values.

[4] For the generalized Pareto distribution we estimate shape and scalar parameters by means of maximum likelihood. See Supplement S5.1 for more details.

[5] Alternative threshold configurations (e.g., quantile levels or anomaly probabilities) can be explored to assess their impact on false positive rates.

Stationarity is essential for accurately modeling the tail behavior of a distribution, as it assumes the probability of extreme events remains constant over time. Independence is critical for the asymptotic convergence to the generalized Pareto distribution, thereby ensuring the validity of maximum likelihood estimation (MLE) and mitigating the risk of modeling clusters of dependent extremes.

To assess the assumption of stationarity in our applications, we applied the Bayesian change-point detection method proposed by Zhao et al. (2019) to the trend component of the Mahalanobis distances for both benchmarks and case study datasets (see Supplementary Material S13. This analysis estimates the probabilities of structural changes in the time series. The results indicate potential change points in both datasets, but with low associated probabilities, suggesting stationarity of the Mahalanobis distance. To further evaluate the robustness of the Mahalanobis distance, we conducted additional experiments involving simulated data where variables are not stationary (random walk with and without drift). Even in these controlled non-stationary scenarios, the change-point probabilities remained consistently low, indicating that the Mahalanobis distance metric is relatively insensitive to non-stationary behavior in the input variables (see Supplementary Material S13).

Despite the aforementioned results, FRA^2Nk can be extended to accommodate scenarios characterized by non-stationary Mahalanobis distance. In such cases, change-point analysis can be used to segment the Mahalanobis distance into stationary intervals. A separate generalized Pareto distribution can then be fitted within each segment, allowing for adaptive threshold estimation based on local distributional properties.

Declustering techniques can be readily integrated into the FRA^2Nk pipeline in contexts of autocorrelated peaks. These involve grouping temporally adjacent observations into clusters and retaining only the maximum value from each cluster for parameter estimation. The resulting set of cluster maxima is then used to fit the generalized Pareto distribution. Notably, in our empirical evaluation across all the considered benchmark datasets, we observed no significant benefit from applying declustering (see Supplementary Material S13). This result is due to the low autocorrelation of peaks in the considered benchmark datasets.

Step 5 - Identifying variables involved in anomalies (to address **R2**)

After detecting anomalies, we try to associate them to specific variables – as to aid domain experts in the investigation of putative causes. We do so by training supervised methods equipped with feature importance techniques, using the detected anomalies in $\hat{y}$ as prediction targets. In more detail, we fit a prediction model on each time interval of interest (i.e., each interval from the test set corresponding to a detected long-lived anomaly), which we combine with arbitrarily selected anomaly-free observations (in our experiments in Sect. 5 we use observations from the final portion of the training set, while in Sect. 7 we use anomaly-free observations from the same interval containing anomalies). We use two well-known prediction models; namely, *random forest* (Breiman 2001) and *logistic regression* (James et al. 2013, ch. 4). Feature importance is evaluated through the *Gini Index* for the former (see Supplementary Material S5.2 and (James et al. 2013, ch. 8)), and through the *Relative Contribution to Deviance Explained* (RCDE) for the latter (see Supplementary Material S5.3). Notably, in Supplement S14, we report SHAP values computed for the

random forest model. This analysis complements the Gini-based feature importance by offering a more robust evaluation of each feature's contribution. The consistency between SHAP and Gini rankings further reinforces confidence in the relevance of the identified variables to the detected anomalies.

As we will see in Sect. 4, our experimental results demonstrate that FRA^2Nk is fast, robust, and accurate across a wide range of tasks. Notably, it outperforms several more complex deep learning and machine learning methods in semi-supervised settings. Motivated by these findings, we have intentionally kept FRA^2Nk simple and interpretable, despite the possibility of incorporating more advanced techniques for specific subtasks.

A practitioner-oriented overview of FRA^2Nk is given in Supplementary Material S3.

4 Evaluation of performance and runtime

We evaluate our procedure in terms of performance and computational cost by comparing it with several state-of-the-art semi-supervised anomaly detection methods on multiple public datasets from the anomaly detection literature.[6] For the selection of methods and datasets to employ in our comparison, we rely on the recent, broad survey by Schmidl et al. (2022).[7] The survey comprised 158 anomaly detection methods from the literature, spanning several domains (see Fig. 3). Of these, 11 are semi-supervised. Here we compare FRA^2Nk to 9 of these semi-supervised methods; namely, LSTM-AD (Malhotra et al. 2015), HealthESN (Chen et al. 2020), Telemanom (Kyle et al. 2018), Random Black Forest (Schmidl et al. 2022), EncDec-AD (Malhotra et al. 2016), DEEPAnT (Munir et al. 2018), Omnianomaly (Su et al. 2019), Robust-PCA (Paffenroth et al. 2018), and Hybrid-KNN (Song et al. 2017) – all run using settings and implementations provided in Schmidl et al. (2022) and in its replicability material at https://github.com/TimeEval/TimeEval-algorithms. We omit LaserDBN (Ogbechie et al. 2017) and TAnoGan (Bashar and Nayak 2020) due to implementation issues, but these were among the bottom ranking in the evaluation by Schmidl et al. (2022). We run FRA^2Nk with both $h = 1$ (no smoothing, FRA^2Nk_1) and $h = 10$ (smoothing, FRA^2Nk_{10}). Since most of the procedures, including ours, provide an *anomaly score* to be compared against a threshold, we homogenize the comparison using the same threshold selection methods, namely MVT and POT, for all procedures. The uniform application of two thresholding techniques (MVT and POT) ensures a fair comparison across methods, particularly since many deep learning models report only anomaly scores. This enables consistent evaluation of all models within a controlled and comparable framework. The only exception is Hybrid-KNN, which returns the probability that an observation is an anomaly. Here we present

[6] Code and replicability material for Sects. 4 and 5 are available on ZENODO at https://zenodo.org/records/15076198.

[7] To avoid selecting competitors ourselves, we prefer to use a recent review that contributes significantly to the literature by evaluating the state of the art in anomaly detection techniques for time series. Furthermore, the absence of a superior procedure forces us to consider all semi-supervised methodologies tested in Schmidl et al. (2022).

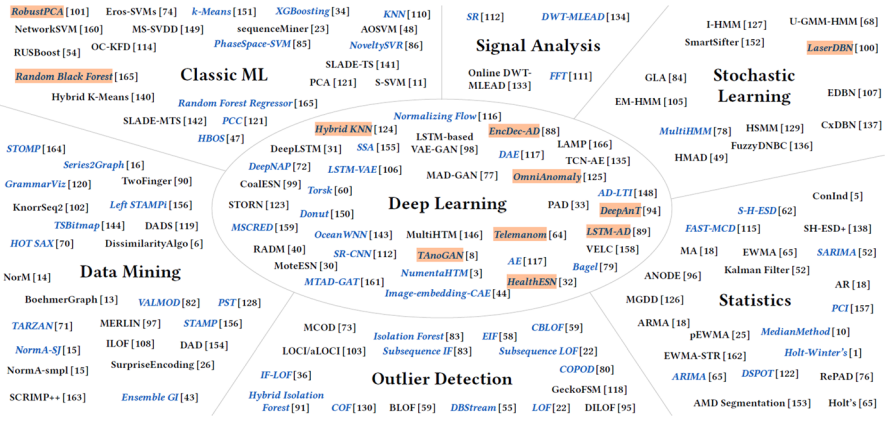

Fig. 3 (Adapted from Schmidl et al. (2022)) The 158 anomaly detection methods for time series data covered in Schmidl et al. (2022), grouped by domain (or *method family*, in the authors' terminology). Citation numbers reported for each method, reported in square brackets, refer to the bibliography in Schmidl et al. (2022). The 11 semi-supervised methods are highlighted in orange

Table 1 Some statistics on the 8 datasets used in our comparison. We consider 4 datasets from each of two collections, listing averages for: number of observations (*Avg. Size*), number of variables (*Avg. Dim.*), and contamination (*Avg. Cont.*) – i.e., the percentage of test observations labeled as anomalies

Collection	Avg. Size	Avg. Dim.	Avg. Cont.
SMD	54 827	38.00	4.04%
Exathlon	94 186	23.25	7.32%

results obtained setting a threshold of 0.8 on such probability; higher thresholds do not change the performance of Hybrid-KNN (see Supplementary Material S6).

The survey by Schmidl et al. (2022) considered 24 collections of public datasets from the anomaly detection literature. Only 2 of these collections contain non-synthetic multivariate time series and comprise anomaly-free data as required by semi-supervised methods; the *Server Machine Datasets*, SMD (Su et al. 2019), and the *Exathlon* (Jacob et al. 2020). From each of these two collections we randomly selected 4 datasets that contain long-lived anomalies. SMD, one of the largest public data repositories available for anomaly detection in multivariate time series, contains 5-week-long datasets collected per minute by a large internet company on 28 machines. It comprises one dataset per machine, each with 38 variables. Exathlon contains 39 datasets with about 23 variables on average. It has been used in the context of industrial fault detection. For both collections, the *training sets* are anomaly-free, while the *test sets* are not and have labels to denote anomalies. The datasets from SMD contain both short- and long-lived anomalies, while the ones from Exathlon contain only long-lived anomalies. Table 1 shows statistics on the datasets.

In terms of performance metrics, we use various well-known indicators of accuracy in anomaly detection; namely: precision, $\text{Prec} = TP/(TP + FP)$; recall, $\text{Recall} = TP/(TP + FN)$; the $F1$ score, $F1 = 2TP/(2TP + FP + FN)$; and the Matthews Correlation Coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

which, ranging in $[-1, 1]$, measures the agreement between true and detected anomalies (Chicco and Jurman 2020). All these metrics combine, in intuitive and effective ways, counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Notably, since they consider individual anomalous observations, they might not be suited to evaluate methods when the focus is on long-lived anomalies. This is because they do not tell us whether long-lived anomalies were flagged by detecting at least one anomaly within them. To capture this, we use an additional metric named *RIC (Ratio of Identified Clusters)*, where a "cluster", i.e., a long-lived anomaly, is counted as identified if at least one observation within its time frame is detected as an anomaly (see Supplementary Material S7).

Table 2 reports, for each method and threshold selection procedure, all performance metrics averaged over the datasets, along with the percentage of datasets where no long-lived anomalies were identified (*%No Anom.*). FRA^2Nk shows consistently high performance. In particular, it always has the best performance with POT, and either the best or second-best with MVT. Smoothing in FRA^2Nk tends to have a positive impact on *F1* and *MCC*, while *%No Anom.* might worsen due to the loss of anomalies in the datasets from SMD.

Next, we focus on performance in detecting long-lived anomalies. For each dataset, we compute the arithmetic mean between *Prec* (which controls false positives) and *RIC* (which targets specifically long-lived anomalies), restricting attention to the portion of the test set that contains long-lived anomalies (see Supplementary Material S8 for more details). In Fig. 4, for each method and thresholding procedure combination, the average of this summary across the considered datasets is plotted against the average runtime of the method – computed as the logarithm of the seconds needed to compute the anomaly scores. In general, we see that POT improves the identification of long-lived anomalies (see also *RIC* results in Table 2; the improvement is most marked for the Telemanom method). Most notably though, with both thresholding procedures and with and without smoothing, FRA^2Nk is located in the upper left corner of Fig. 4, representing the best-performing method with the least runtime. In Supplementary Material S9 we report runtimes separately for training and test sets.

In summary, our comparison demonstrates that FRA^2Nk outperforms competitors in both anomaly detection accuracy and runtime, addressing **R1**. In Sect. 6 and Supplementary Material S10, S11 and S12, we compare FRA^2Nk with the competitor semi-supervised methods on further aspects using simulated data.

5 Evaluation of feature importance

The last step of FRA^2Nk uses random forest or logistic regression to associate detected anomalies to specific variables. Here, we try and evaluate whether and how feature importance assessed by these prediction models can shed light on the putative causes of anomalies. The datasets in SMD comprise information on the variables that caused each anomaly. We thus use the first dataset from SMD for our evaluation,

Table 2 Performance evaluation for each method and threshold selection procedure. In the POT case, HealthESN and RBForest were run on 7 datasets, Telemanom on 6, and Omnianomaly on 5 because the parameters of the generalized Pareto distribution could not be calculated on some of the datasets (see Supplementary Material S5.1). For the Exathlon collection, EncDec-AD failed to calculate anomaly scores. In addition to the performance metrics, which are averaged over the datasets (standard deviations are reported in parentheses), the last column provides the percentage of considered datasets where no long-lived anomalies were identified (*%No Anom.*). The best performance for each metric and thresholding procedure is in bold

Method	*Prec*	*Recall*	*F1*	*MCC*	*RIC*	*%No Anom.*
MVT						
LSTM-AD	0.667 (0.37)	0.124 (0.20)	0.104 (0.15)	0.168 (0.15)	0.309 (0.44)	62.5%
HealthESN	0.582 (0.39)	0.511 (0.51)	0.211 (0.16)	0.209 (0.13)	**0.928** (0.14)	**0.0%**
Telemanom	0.713 (0.17)	0.054 (0.08)	0.104 (0.14)	0.148 (0.21)	0.229 (0.37)	62.5%
RBForest	0.998 (0.01)	0.014 (0.02)	0.030 (0.04)	0.140 (0.07)	0.365 (0.45)	50.0%
EncDec-AD	0.370 (0.52)	0.003 (0.01)	0.009 (0.02)	−0.007 (0.10)	0.025 (0.07)	50.0%
DeepAnT	0.827 (0.23)	0.100 (0.16)	0.140 (0.22)	0.278 (0.22)	0.246 (0.27)	62.5%
OmniAnomaly	0.907 (0.09)	0.093 (0.14)	0.152 (0.22)	0.318 (0.26)	0.272 (0.38)	62.5%
RobustPCA	**1.000** (0.00)	0.033 (0.05)	0.061 (0.08)	0.222 (0.11)	0.192 (0.28)	50.0%
FRA^2Nk_1	**1.000** (0.00)	0.304 (0.43)	0.342 (0.44)	0.427 (0.43)	0.813 (0.37)	12.5%
FRA^2Nk_{10}	0.872 (0.31)	0.491 (0.43)	**0.650** (0.41)	**0.630** (0.45)	0.823 (0.35)	12.5%
POT						
LSTM-AD	0.603 (0.37)	0.155 (0.25)	0.117 (0.14)	0.160 (0.15)	0.376 (0.47)	62.5%
HealthESN	0.529 (0.37)	0.496 (0.47)	0.244 (0.12)	0.219 (0.15)	**1.000** (0.00)	**0.0%**
Telemanom	0.852 (0.17)	0.460 (0.41)	0.522 (0.38)	0.482 (0.41)	0.638 (0.43)	16.67%
RBForest	0.614 (0.28)	0.533 (0.38)	0.421 (0.28)	0.369 (0.31)	**1.000** (0.00)	**0.0%**
EncDec-AD	0.573 (0.52)	0.018 (0.02)	0.038 (0.04)	0.035 (0.14)	0.112 (0.13)	25.0%
DeepAnT	0.715 (0.32)	0.266 (0.32)	0.140 (0.34)	0.380 (0.37)	0.469 (0.51)	50.0%
OmniAnomaly	0.883 (0.12)	0.164 (0.17)	0.315 (0.23)	0.320 (0.26)	0.436 (0.41)	40.0%
RobustPCA	0.890 (0.09)	0.061 (0.08)	0.110 (0.15)	0.250 (0.14)	0.475 (0.51)	50.0%
FRA^2Nk_1	**0.901** (0.17)	0.585 (0.44)	0.635 (0.39)	0.624 (0.41)	**1.000** (0.00)	**0.0%**
FRA^2Nk_{10}	0.841 (0.25)	0.673 (0.42)	**0.666** (0.38)	**0.645** (0.40)	0.844 (0.35)	12.5%
$\Pr > 0.8$						
Hybrid KNN	0.309 (0.22)	**0.761** (0.21)	0.312 (0.43)	0.143 (0.31)	0.959 (0.22)	0.0%

focusing in particular on 5 anomalies, among the 8 within it, that are long-lived (they last several minutes, see Supplementary Material S8).

The first four steps of FRA^2Nk run with $h = 1$ and POT detect all these long-lived anomalies. We run step 5 by training random forest and logistic regression for each such anomaly, using 1 000 observations from the test set containing the anomaly, and 1 000 additional anomaly-free observations from the training set (see Supplementary Material S8).

Table 3 summarizes results. The *real* causes indicated by the dataset description are reported in the column labeled *Causes*; variables here are arbitrarily numbered based on the column order in the data set, and listed based on such numbers – but SMD does not provide information on the importance of the causes; that is, the variables are not ranked. Sections *Ranked by RF* and *Ranked by LR* report the variables identified as most relevant using random forest and logistic regression, respectively, ranked by decreasing importance (see Supplementary Material S14). For each anom-

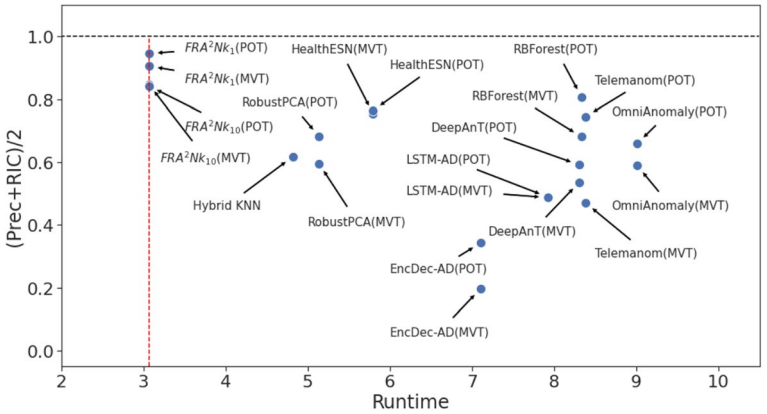

Fig. 4 Performance of each method and threshold procedure in detecting long-lived anomalies (y-axis), vs runtime expressed as logarithms of seconds needed to compute anomaly scores (x-axis). Performances and runtimes are averaged across the considered datasets. Hybrid KNN uses the threshold $\Pr > 0.8$

aly a, we include the $v(a)$ most important variables, $v(a)$ being the number of variables indicated as causes in the dataset description. The results of step 5 are indeed coherent with the causes as provided by the dataset description; 82.1% and 79.1% of the variables identified by our procedure using *RF* and *LR*, respectively, are among such causes. Supplement S14 presents the SHAP values computed for the random forest. While the Gini Index provides a global ranking of features based on their average contribution to impurity reduction, SHAP values offer a complementary, theoretically grounded perspective by quantifying the marginal contribution of each feature to individual predictions. Therefore, consistency between Gini and SHAP rankings increases our confidence that the identified variables are relevant to the anomalies. Results show that the SHAP value analysis corroborates the feature importance rankings derived from the Gini index. These results demonstrate how FRA^2Nk can effectively address **R2**.

6 Simulation experiments

In this section, we investigate the effectiveness of FRA^2Nk in scenarios with increased temporal dependence, and where the training data are (erroneously) contaminated with anomalies, i.e., where Assumption 1 does not hold.

For all scenarios, the initial dataset is composed of 40,000 training observations, 10,000 test observations, and 30 independent standard Gaussian variables. Anomalies are introduced into the test set for the first five variables by adding 10 short-lived anomalies (lasting 1 observation, each every 1000 observations), and one long-lived anomaly (lasting 750 observations). They are generated from a $N(30, 1)$ or a $N(-30, 1)$. For each simulation setting, we generate three datasets and report the average of the considered metrics.

Table 3 Variables labeled as causes in the dataset vs variables ranked as important for the anomalies by random forest (*RF*) and logistic regression (*LR*). We mark such important variables in bold if they are among the labeled causes. With very few exceptions (e.g., anomaly 4, LR) the top portion of the rankings is occupied by variables labeled as causes

Anom.	*Causes*	*Ranked by RF,* $h = 1$	*Ranked by LR,* $h = 1$
1	1,9,10,12,13,14,15	**10**,**12**,**13**,**9**,**15**,**14**,**1**	**10**,**15**,**9**, *6*,**13**,*2*,**1**
2	1,2,3,4,6,7,9,10,11,12	**34**,**20**,**35**,**1**,**36**,**13**,**10**,**28**,**29**,**31**	**10**,**1**,**6**,**4**,**9**,**15**,*23*,**2**,**11**,**36**
	13,14,15,16,19,20,21,22,24,25	**12**,**22**,**2**,**21**,**3**,**33**,**25**,**19**,**15**,**14**	**35**,**3**,**32**,**16**,**21**,**20**,**19**,**13**,**34**,**14**
	26,27,28,29,30,31,32,33,34,35,36	**6**,**26**,**4**,**32**,*23*,**16**,**7**,**11**,**24**,**30**	**24**,**31**,**26**,**29**,**33**,**25**,**22**,**30**,**28**,**12**,**7**
3	1,2,9,10,12,13,14,15	**10**,**12**,**13**,*20*,**1**,**9**,*28*,*22*	**10**,**15**,**1**,**13**,*34*,*29*,**9**,**14**
4	1,2,3,4,9,10,11,12,13,14,15,16,25,28	**10**,**12**,**13**,*7*,*6*,*20*,*34*,**9**,*29*,*36*,**25**,**28**,*30*,**15**	**10**,*6*,*29*,**14**,**11**,*23*,**16**,**13**,*34*,*33*,**2**,**9**,**1**,**15**
5	1,9,10,12,13,14,15	**10**,**13**,**12**,**9**,**15**,**14**,*23*	**10**,**9**,*33*,**15**,*23*,*24*,*26*

6.1 Temporal dependence

Data are generated from the process $x_{i,t} = \phi x_{,it-1} + u_{i,t}$. For the variables not causing anomalies ($i = 6, \ldots, 30$) we draw $u_{i,t} \sim N(0, 1)$ for all observations $t = 1, \ldots, 50000$. For the variables causing anomalies ($i = 1, \ldots, 5$) we draw $u_{i,t} \sim N(0, 1)$ if t is not involved in an anomaly, and $u_{i,t} \sim N(30, 1)$ for $i = 1, 2, 3$, or $u_{i,t} \sim N(-30, 1)$ for $i = 4, 5$, if t is involved in an anomaly. We consider four values of ϕ: 0.0 (in the initial dataset), 0.3, 0.6, and 0.9. For $\phi = 0$ short-lived anomalies last 1 observation, while for $\phi > 0$ they last longer due to temporal dependence.

Results are reported in columns (a) of Table 4. Notably, temporal dependence affects precision, while it does not affect recall. All methods decrease in precision as ϕ increases; however, FRA^2Nk_1 and FRA^2Nk_{10} either outperform or perform on par with the competitors for all values of ϕ. Moreover, FRA^2Nk can be "tuned" to have better precision also under extreme temporal dependence. In the dataset obtained by setting $\phi = 0.9$, the 10 short-lived anomalies have longer duration – but employing higher values for the smoothing parameter ($h = 100$), FRA^2Nk_{100} recovers a higher precision. To further study the impact of temporal dependence on long-lived anomalies, we remove the 10 short-lived ones. Results are reported in columns (b) of Table 4. Again, recall is unaffected by temporal dependence, but here all methods maintain high precision as temporal dependence increases.

6.2 Training set contamination

Next, we study the impact of anomalies in the training data (failures of Assumption 1). To do this, we add long-lived anomalies as to occupy $100 \times \alpha\%$ of the training set. The anomalies are generated as described in Sect. 6.1, and we consider four contamination levels: $\alpha = 0$ (initial dataset), 0.03, 0.06, and 0.09. Results are reported in columns (c) of Table 4.

We focus exclusively on the recall metric because, intuitively, the presence of anomalies in the training set can cause methods to misclassify true anomalies in the test set as normal behavior, leading to the masking effect (see Sect. 2). This phenomenon leads to a drop in recall and is particularly problematic in the context of semi-supervised anomaly detection, where the training data is assumed to be anomaly-free. Our results show a consistent pattern across all methods—including FRA^2Nk—and across all thresholding approaches (MVT, POT, and the probability threshold used in Hybrid KNN). Specifically, when long-lived anomalies are present in the training set, all methods experience a substantial decline in recall, indicating a rise in false negatives (masking effect).

Additional experiments are reported in Supplementary Material S10 and S11, where we assess robustness to non-elliptical data and scalability. Notably, FRA^2Nk maintained high precision and recall on data from Dirichlet and Beta-Gamma mixtures, indicating robustness to deviations from elliptical distributions. In terms of scalability, runtime increases more with the size of the training set than the test set, due to the cost of estimating the mean and covariance matrix. Finally, Supplementary Material S12 reports the precision results under training set contamination. As

Table 4 Average of precision and recall of FRA^2Nk and competitor methods obtained on three simulated datasets. Columns *Init.* provide precision and recall computed on the initial dataset ($\phi = 0$, and $\alpha = 0$). Columns (a) report the impact of temporal dependence (ϕ) in the case of short- and long-lived anomalies, while columns (b) only for long-lived ones. Columns (c) report the impact of training sample contamination (long-lived anomalies covering different percentages of the training data)

Method	*Precision*							*Recall*									
	Init.	(a)			(b)			*Init.*	(a)			(b)			(c)		
		ϕ			ϕ				ϕ			ϕ			α		
		0.3	0.6	0.9	0.3	0.6	0.9		0.3	0.6	0.9	0.3	0.6	0.9	0.03	0.06	0.09
	MVT																
LSTM-AD	0.69	0.69	0.67	0.54	0.95	0.96	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.04	0.04	0.03
HealthESN	1.00	0.99	0.95	0.72	1.00	0.99	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.01	0.01
Telemanom	0.34	0.26	0.26	0.17	0.32	0.26	0.37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.11	0.12	0.10
RBForest	1.00	0.98	0.94	0.70	0.99	0.99	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.01	0.01
EncDec-AD	0.48	0.47	0.46	0.41	0.90	0.90	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.04	0.01
DeepAnT	0.61	0.60	0.60	0.53	0.94	0.94	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
OmniAnomaly	0.41	0.41	0.41	0.37	0.88	0.88	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.05	0.05
RobustPCA	1.00	0.99	0.95	0.73	1.00	0.99	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
FRA^2Nk_1	1.00	0.99	0.96	0.74	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
FRA^2Nk_{10}	1.00	1.00	0.94	0.72	1.00	0.99	0.96	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
FRA^2Nk_{100}	0.94	0.94	0.96	0.85	0.95	0.97	0.93	0.99	0.99	0.99	0.99	1.00	1.00	1.00	0.02	0.01	0.00
	POT																
LSTM-AD	0.66	0.67	0.64	0.52	0.94	0.87	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.05	0.07	0.05
HealthESN	0.98	0.97	0.94	0.71	0.98	0.98	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.02	0.01	0.02
Telemanom	0.40	0.17	0.19	0.14	0.39	0.16	0.27	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.13	0.16	0.16
RBForest	0.91	0.89	0.85	0.52	0.91	0.92	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.02	0.02	0.02
EncDec-AD	0.46	0.47	0.46	0.41	0.89	0.89	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.02	0.00	0.03
DeepAnT	0.60	0.58	0.59	0.52	0.93	0.86	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.01	0.01
OmniAnomaly	0.41	0.41	0.40	0.36	0.88	0.86	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.05	0.05
RobustPCA	0.99	0.97	0.94	0.71	0.98	0.98	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.02	0.01
FRA^2Nk_1	0.99	0.98	0.94	0.72	0.98	0.98	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.02	0.01

Table 4 (continued)

Method	*Precision*							*Recall*									
	Init.	(a)			(b)			*Init.*	(a)			(b)			(c)		
		ϕ			ϕ				ϕ			ϕ			α		
		0.3	0.6	0.9	0.3	0.6	0.9		0.3	0.6	0.9	0.3	0.6	0.9	0.03	0.06	0.09
	MVT																
FRA^2Nk_{10}	0.99	0.99	0.92	0.70	0.99	0.98	0.94	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.06	0.03	0.01
FRA^2Nk_{100}	0.91	0.91	0.92	0.78	0.92	0.93	0.89	0.99	0.99	0.99	0.99	1.00	1.00	1.00	0.12	0.07	0.02
	$\Pr > 0.8$																
Hybrid KNN	0.73	0.51	0.47	0.31	0.55	0.50	0.37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.04	0.02

expected, all methods exhibit a decline in performance; however, FRA^2Nk consistently outperforms the competing approaches.

7 Case study

In this section, we demonstrate the effectiveness of FRA^2Nk on an industrial case study involving a tissue machine of the A.Celli company, leader in the supply of machinery and technologies for the paper and nonwoven products industry. This application was performed in close collaboration with domain experts who validated its results. The data consisted of 70 000 observations collected per second on 217 variables. We used the first 60,000 for training, and the last 10,000 for testing. In the considered time interval, the machine handled only one type of product, allowing us to restrict attention to 119 variables deemed relevant for this one product. For the smoothing in step 1 of FRA^2Nk, we considered two window sizes; $h = 1$ (no smoothing), and $h = 10$. For the feature importance assessment, we focused on random forest (*RF* performed better than *LR* in Sect. 5).

We start from the analysis with no smoothing ($h = 1$). In step 2, because of high multicollinearity in our case study, VIFs reduced the variables from 119 to 60. In step 3 we computed the anomaly scores, and in step 4 we thresholded them with MVT to detect the four anomalies shown in Fig. 5 (a-b). Zooming in (Fig. 5 b), we can see that three are short-lived and one is long-lived. However, given the fact that they are very close, the domain experts interpreted the 'cluster' of four detected anomalies as a single long-lived anomaly. Therefore, in step 5 we investigated the putative causes of the cluster of detected anomalies as a whole. We did this by training a random forest on the entire interval (2000 observations) shown in Fig. 5 (b). Differently from Sect. 5, here we do not need to add anomaly-free observations from the training sample, because such interval contains plenty of anomaly-free observations. The domain experts deemed the top $v = 5$ variables as sufficient and relevant for interpreting the causes of the anomalies. These are: *WERecirculationFanPower*, *DERecirculationFanPower* (two energy expenditure variables for the fans that manage humidity and the drying process of the sheet), *MCC1PowerConsumption* (power consumption of the electrical distribution panel), *DEGasConsumption* (gas consumption of the drying process), and *V6-Speed* (speed of a fan used to dry the sheet). Figure 5 (c), which depicts the time series of these 5 variables, suggests the presence of a long-

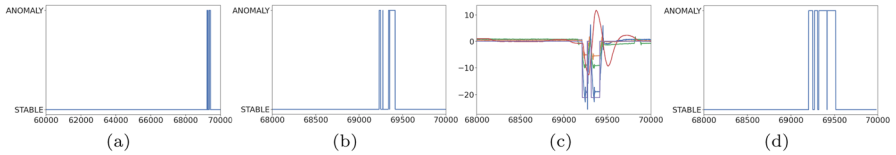

Fig. 5 Case study results using MVT thresholding. All x-axes contain the time-frame of interest. Panel (**a**) shows anomalies detected in the test set by FRA^2Nk_1 (no smoothing, $h = 1$). Panel (**b**) highlights such anomalies zooming in on the last 2000 observations. Panel (**c**) shows the values of the standardized time series (y-axis) of the 5 most important variables as identified by a random forest. Panel (**d**) is the same as Panel (**b**), but for FRA^2Nk_{10} (smoothing with $h = 10$)

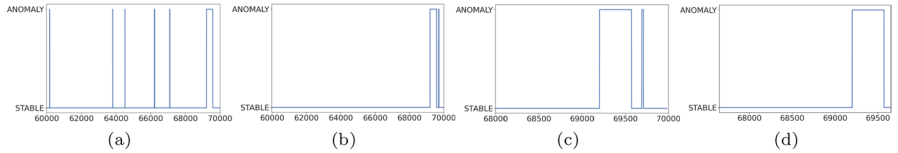

Fig. 6 Results in the case of POT. Panels (**a**) and (**b**) show the anomalies detected in the test set by FRA^2Nk_1 and FRA^2Nk_{10}, respectively. Panel (**c**) highlights the anomalies detected by FRA^2Nk_{10} by showing the last 2000 observations only. Panel (**d**) highlights the long-lived anomaly detected by FRA^2Nk_{10}

lived anomaly. Indeed, the domain experts were able to validate our findings as an actual process anomaly caused by overheating of the system during the paper drying phase. Running FRA^2Nk with smoothing ($h = 10$) lends further support to the existence of one long-lived anomaly. The VIFs in step 2 leads to the same 60 variables obtained without smoothing. Thresholding the resulting anomaly scores, again with MVT, detects one, essentially uninterrupted long-lived anomaly in the same time interval (Fig. 5 d). A random forest, trained on smoothed data, ranks as top 5 variables *BRConsistency* (pulp consistency setpoint), *WERecirculationFanPower*, *MCC1PowerConsumption*, *Thermocompressor* (sensor on drying phase), and *SF-Refiner-Outlet-Pressure* (outlet pressure of the SF refiner). While three of these variables differ from those identified using $h = 1$, the domain experts point to the same cause, i.e., overheating during the paper drying phase.

Next, we switched from MVT to POT thresholding. Figure 6 (a) displays results without smoothing ($h = 1$). POT detects many more anomalies than MVT (Fig. 5 (a)), and especially many short-lived ones attributed by domain experts to sensor noise. Figures 6 (b)-(c) display results with smoothing ($h = 10$), which makes FRA^2Nk more suited to long-lived anomalies; FRA^2Nk_{10} detects 1 long-lived and 1 short-lived anomaly, discarding other sensor noise detected by FRA^2Nk_1. The domain experts deemed the long-lived anomaly to be the same as that in Fig. 5 (d). Indeed, the time interval involved is almost identical. Training a random forest on the 2000 observations in Fig. 6 (d) we find that 3 of the 5 top ranked variables are shared with those identified by FRA^2Nk_{10} with MVT thresholding; namely, *BRConsistency*, *WERecirculationFanPower* and *MCC1PowerConsumption*. The other top variables are *SheetON* (an indicator of whether the paper passes from the last roll to the winder), and *DERecirculationFanPower* (also identified by FRA^2Nk_1 with MVT). According to the domain experts, step 5 run for the short-lived anomaly in Fig. 6 (c) produces a variables ranking indicating a system overheating (see Supplementary Material S15).

This analysis exemplifies an important aspect of anomaly detection: the *real* cause underlying an anomaly might not be directly expressed by the observed variables; domain knowledge is required to reconstruct causes. In our case study, the variable "system overheating" does not exist. However, the variables selected by step 5 of FRA^2Nk with various specifications (with or without smoothing, with different thresholding) all point to the slowing and cooling of the production system, allowing our industrial partners to identify system overheating as the underlying cause.

8 Concluding remarks

In this article we introduced FRA^2Nk, a semi-supervised method for anomaly detection. We use the term *semi-supervised* as in Schmidl et al. (2022), to mean a method that is trained on anomaly-free data. The main message of our work is that a procedure built combining relatively simple statistical techniques (our FRA^2Nk) can outperform state-of-the-art methods in its ability to detect anomalies, runtime, and robustness to a number of potential problems in the data. We target industrial processes, characterized by multicollinear time series with unknown distributions. Notably, the choice of statistical tools employed in FRA^2Nk was driven by the dual objective of optimizing efficiency (precision, runtime, etc) while maintaining simplicity, the two hallmarks of our approach. Of course, as discussed in Sect. 3, there are alternative (more complex) techniques that could be utilized in FRA^2Nk.

Nevertheless, we have decided to keep FRA^2Nk simple, as it outperforms state-of-the-art methods on benchmark datasets from the literature (both in terms of performance and runtime). Furthermore, our experiments show that it meets the flexibility, reliability and simplicity requirements recently highlighted, e.g., in Schmidl et al. (2022). In fact, FRA^2Nk outperformed competing methods in a broad variety of datasets and domains (flexibility) and it successfully discovered anomalies in all such datasets (reliability). Moreover, employing straightforward tools, FRA^2Nk does not require much parameter tuning (simplicity), apart for the choice of the smoothing parameter h. The value of such parameter is critical, however in all experiments FRA^2Nk demonstrated high performance with 3 preselected values of h (1, 10 and 100).

Importantly, FRA^2Nk also allows users to identify variables relevant for the discovered anomalies. Thus, FRA^2Nk meets several key requirements for anomaly detection in industrial settings; namely, to enable domain experts to quickly and accurately identify time intervals when anomalies occur, and to help decipher their putative causes.

We envision a number of avenues for future work. First, given its simplicity and low computational demands, FRA^2Nk may be particularly well-suited for use on very large or streaming datasets. Second, FRA^2Nk could be extended with the ability to rank anomalies by their impact on domain-specific key performance indicators (KPIs). Third, as discussed in Sect. 3, FRA^2Nk could be extended to better accommodate highly non-stationary and temporal dependence settings by incorporating change-point detection and declustering. Fourth, although our current configuration demonstrates strong precision performance, further parameter tuning could enhance robustness in future applications. In particular, we plan to investigate alternative threshold settings—such as varying the quantile levels used to define peaks and the probabilities assigned to classify them as anomalies—with the goal of identifying configurations that further improve false discovery rate control. Finally, Step 5 could be further improved including techniques to counteract the effects of unbalanced counts of anomalous and anomaly-free observations in the training of prediction models (random forests or logistic regression, (see, e.g., Fithian and Hastie (2014)).

In the case study illustrated in Sect. 7 the time intervals used to train random forests contained more anomaly-free than anomalous observations, but the variables identified as relevant in Step 5 were validated by domain experts. Therefore, FRA^2Nk appeared not to be affected by the unbalanced training in this instance – though this might not hold in general.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-026-00667-8.

Acknowledgements The authors would like to thank the A.Celli Nonwovens S.p.A. company and the Extreme Automation group for providing the industrial data and technical support that made this study possible.The authors also thank the journal reviewers for their valuable comments, which helped improve the quality of the manuscript.

Funding Open access funding provided by Scuola Superiore Sant'Anna within the CRUI-CARE Agreement. This work was supported by the Fsc regional Tuscan projects AUTOXAI2 J53D21003810008 and AISLEA2 J54D23000780005, and by project the SMaRT COnSTRUCT project (CUP J53C24001460006), in the context of FAIR (PE0000013, CUP B53C22003630006) under the Italian National Recovery and Resilience Plan (Mission 4, Component 2, Line of Investment 1.3) funded by the European Union - NextGenerationEU.

Data availability Data and codes to replicate the results reported in this paper are available on Zenodo at https://zenodo.org/records/15076198.

Declarations

Ethics approval and consent to participate This study did not require ethics approval or informed consent, as no human or animal subjects were involved and no personal data were used.

Competing interests 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

Arias-Castro E, Castro RM, Tánczos E, Wang M (2018) Distribution-free detection of structured anomalies: Permutation and rank-based scans. J Am Stat Assoc 113(522):789–801

Balkema AA, De Haan L (1974) Residual life time at great age. Ann Probab 2(5):792–804

Bellini T (2016) The forward search interactive outlier detection in cointegrated var analysis. Adv Data Anal Classif 10(3):351–373

Blázquez-García A, Conde A, Mori U, Lozano JA (2021) A review on outlier/anomaly detection in time series data. ACM Comput Surv 54(3):1–33

Bashar MA, Nayak R (2020) Tanogan: Time series anomaly detection with generative adversarial networks. In: 2020 IEEE SSCI, pp. 1778–1785. IEEE
Breiman L (2001) Random forests. Mach Learn 45(1):5 32
Boudt K, Rousseeuw PJ, Vanduffel S, Verdonck T (2020) The minimum regularized covariance determinant estimator. Stat Comput 30(1):113–128
Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 21(1):1–13
Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8(1):5–26
Craney TA, Surles JG (2002) Model-dependent variance inflation factor cutoff values. Qual Eng 14(3):391–403. https://doi.org/10.1081/QEN-120001878
Choi K, Yi J, Park C, Yoon S (2021) Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access 9:120043–120065
Chen Q, Zhang A, Huang T, He Q, Song Y (2020) Imbalanced dataset-based echo state networks for anomaly detection. Neural Comput Appl 32:3685–3694
Deraemaeker A, Worden K (2018) A comparison of linear approaches to filter out environmental effects in structural health monitoring. MSSP 105:1–15. https://doi.org/10.1016/j.ymssp.2017.11.045
Fisch AT, Bardwell L, Eckley IA (2022) Real time anomaly detection and categorisation. Stat Comput 32(4):55
Fithian W, Hastie T (2014) Local case-control sampling: Efficient subsampling in imbalanced data sets. Ann Stat 42(5):1693
Fibbi E, Perrotta D, Torti F, Van Aelst S, Verdonck T (2024) Co-clustering contaminated data: a robust model-based approach. Adv Data Anal Classif 18(1):121–161
Grosskopf M, Myers K, Lawrence E, Bingham D (2022) Temporal characterization and filtering of sensor data to support anomaly detection. Technometrics 64(4):475–486. https://doi.org/10.1080/00401706.2022.2124308
Gnedenko B (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of mathematics 44(3):423–453
Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton, NJ, p 799
Hubert M, Debruyne M (2010) Minimum covariance determinant. Wiley interdisciplinary reviews Computational statistics 2(1):36–43
Huang J, Sun X, Yang X, Peng K (2022) Fault detection for chemical processes based on non-stationarity sensitive cointegration analysis. ISA Transactions 129:321–333. https://doi.org/10.1016/j.isatra.2022.02.010
Hubert M, Veeken S (2008) Outlier detection for skewed data. J Chemom 22(3–4):235–246. https://doi.org/10.1002/cem.1123
Jacob V, Song F, Stiegler A, Rad B, Diao Y, Tatbul N (2020) Exathlon: A benchmark for explainable anomaly detection over time series. arXiv:2010.05073
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R Vol. 103. New York: springer
Kalina J, Tichavskỳ J (2022) The minimum weighted covariance determinant estimator for high-dimensional data. Adv Data Anal Classif 16(4):977–999
Kyle H, Valentino C, Christopher L, Ian C, Tom S (2018) Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: KDD, pp. 387–395
Kim E, Yamada Y, Okamoto S (2022) Detection error contaminated by outliers to classify density profiles dependent on the relative speed between a mimo sensor and a human hand. IEEE Access 10:90576–90585. https://doi.org/10.1109/ACCESS.2022.3201563
Leadbetter MR (1974) On extreme values in stationary sequences. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 28:289–303
Loynes RM (1965) Extreme values in uniformly mixing stationary stochastic processes. The Annals of Mathematical Statistics 36(3):993–999
Mahalanobis PC (1936) On the generalized distance in statistics. NISI 80:S1–S7
Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M (2019) Robust Statistics: Theory and Methods (with R), John Wiley & Sons
Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G (2016) Lstm-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148
Munir M, Siddiqui SA, Dengel A, Ahmed S (2018) Deepant: A deep learning approach for unsupervised anomaly detection in time series. IEEE Access 7:1991–2005

Malhotra P, Vig L, Shroff G, Agarwal P et al (2015) Long short term memory networks for anomaly detection in time series. Esann 2015:89
Noble WS (2009) How does multiple testing correction work Nat Biotechnol 27(12):1135–1137
Ogbechie A, Díaz-Rozo J, Larrañaga P, Bielza C (2017) Dynamic bayesian network-based anomaly detection for in-process visual inspection of laser surface heat treatment. In: Machine Learning for Cyber Physical Systems, pp. 17–24. Springer
Penny KI (1996) Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. J. R. Stat.: Series C (Appl. Statistics) 45(1):73–81
Pickands J III (1975) Statistical inference using extreme order statistics. Annals Statistics 3:119–131
Paffenroth R, Kay K, Servi L (2018) Robust PCA for anomaly detection in cyber networks. arXiv preprint arXiv:1801.01571
Raymaekers J, Rousseeuw PJ (2021) Fast robust correlation for high-dimensional data. Technometrics 63(2):184–198. https://doi.org/10.1080/00401706.2019.1677270
Siffer A, Fouque P-A, Termier A, Largouet C (2017) Anomaly detection in streams with extreme value theory. In: KDD'17, pp. 1067–1075. https://doi.org/10.1145/3097983.3098144
Song H, Jiang Z, Men A, Yang B et al (2017) A hybrid semi-supervised anomaly detection model for high-dimensional data. Computational intelligence and neuroscience 2017:8501683
Schmidl S, Wenig P, Papenbrock T (2022) Anomaly detection in time series: A comprehensive evaluation. Proc. VLDB Endow. 15(9):1779–1797. https://doi.org/10.14778/3538598.3538602
Su Y, Zhao Y, Niu C, Liu R, Sun W, Pei D (2019) Robust anomaly detection for multivariate time series through stochastic RNN. In: KDD'19 pp. 2828–2837. https://doi.org/10.1145/3292500.3330672
Tonini S, Barsacchi F, Chiaromonte F, Licari D, Vandin A (2023) Towards novel statistical methods for anomaly detection in industrial processes. In: Companion of the 2023 ACM/SPEC International Conference on Performance Engineering. ICPE '23 Companion, pp. 147–153. Association for Computing Machinery, New York, NY, USA https://doi.org/10.1145/3578245.3585036
Tomarchio SD, Gallaugher MP (2024) Mixtures of regressions using matrix-variate heavy-tailed distributions. Adv Data Anal Classif. https://doi.org/10.1007/s11634-024-00585-7
Talagala PD, Hyndman RJ, Smith-Miles K (2021) Anomaly detection in high-dimensional data. J Comput Graph Stat 30(2):360–374
Tiku ML, Islam MQ, Qumsiyeh SB (2010) Mahalanobis distance under non-normality. Statistics 44(3):275–290. https://doi.org/10.1080/02331880903043223
Vinue G, Epifanio I (2021) Robust archetypoids for anomaly detection in big functional data. Adv Data Anal Classif 15(2):437–462
Zeller CB, Cabral CRB, Lachos VH, Benites L (2019) Finite mixture of regression models for censored data based on scale mixtures of normal distributions. Adv Data Anal Classif 13(1):89–116
Zhao K, Wulder MA, Hu T, Bright R, Wu Q, Qin H, Li Y, Toman E, Mallick B, Zhang X et al (2019) Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A bayesian ensemble algorithm. Remote Sens Environ 232:111181

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Simone Tonini[1] · Andrea Vandin[1,2] · Francesca Chiaromonte[1,3] · Daniele Licari[1] · Fernando Barsacchi[4]

✉ Andrea Vandin
andrea.vandin@santannapisa.it

Simone Tonini
simone.tonini@santannapisa.it

Francesca Chiaromonte
fxc11@psu.edu

Daniele Licari
daniele.licari@santannapisa.it

Fernando Barsacchi
f.barsacchi@acelli.it

[1] L'EMbeDS & Inst. of Economics, Sant'Anna School for Advanced Studies, Piazza Martiri della Libertà, 33, 56127 Pisa, Italy

[2] DTU Technical University of Denmark, Anker Engelunds Vej 101 Kongens Lyngby, 2800 Lyngby, Denmark

[3] The Pennsylvania State University, 201 Old Main, University Park, 16802 Centre County, US

[4] A.Celli, via Romana Ovest 252, Porcari, 55016 Lucca, Italy