Computers & Industrial Engineering 190 (2024) 110064

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Variable contribution analysis in multivariate process monitoring using permutation entropy

Praise Otito Obanya [a,f,*], Roelof L.J. Coetzer [b,c], Carel Petrus Olivier [d], Tanja Verster [e,f]

[a] Unit for Data Science and Computing, North-West University, Potchefstroom, 2531, South Africa
[b] School of Industrial Engineering, North-West University, Potchefstroom, 2531, South Africa
[c] Focus Area for Pure and Applied Analytics, North-West University, Potchefstroom, 2531, South Africa
[d] Pure and Applied Analytics, Department of Mathematics and Applied Mathematics, North-West University, Mafikeng, 2745, South Africa
[e] Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, 2531, South Africa
[f] National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch, 7600, South Africa

ARTICLE INFO

Keywords:
Multivariate statistics
Permutation entropy
Process monitoring
Tennessee Eastman process
Variable contributions

ABSTRACT

Multivariate statistical process monitoring is widely used in industrial processes for performance monitoring and fault detection. Once a fault has been detected, fault diagnoses must be performed to identify the variables that are responsible for the fault or the deviation from normal operation. In this paper, we present a new data-driven method for variable contribution analysis to a fault or abnormal behaviour in a process. The method is based on permutation entropy (PE), which measures the order of fluctuations within a time series. The PE method for variable contribution analysis is independent of fault detection statistics, and is used to evaluate which variables experienced the greatest change in their expected behaviour over time, immediately after the fault has been identified. We show how PE can be used to identify and interpret the variables responsible for, or affected by faults in an industrial process. The well-known Tennessee Eastman Process (TEP) is used to illustrate the application of PE for variable contribution analysis. The results show that PE is an efficient analysis tool for specifying variable contributions to faults.

1. Introduction

The advent of the Fourth Industrial Revolution (4IR) presents industries with great opportunities for performance enhancements owing to data availability, massive technological advancements and enhanced analytical platforms (Reis & Gins, 2017). Due to the industrial internet of things (IIoT) and extensive use of manufacturing execution systems (MES), more data are generated every second across value chains of complex processing and manufacturing plants. The large volumes of data becoming available must be utilized for real-time process monitoring and diagnoses to ensure longer term performance sustainability, process health and conformance to environmental compliance. Hence, industries have inaugurated the utilization of digitalization and systems integration towards improving manufacturing processes, including real-time quality process monitoring and control and predictive maintenance (Farahani et al., 2019). Ji and Sun (2022) mentioned two steps that are required in the implementation of process monitoring — fault detection, which aims at ascertaining whether or not a process is performing in accordance to normal conditions, and fault diagnosis, which is aimed at determining the root cause of a fault. De Lázaro et al. (2015) classified the tools designed for fault detection and diagnosis into three categories — quantitative models based on analytical approaches, qualitative models based on knowledge such as human expertise and data-driven models, which rely on the historical record of the process.

Multivariate statistical methods and data-driven methods are widely used for process monitoring and fault detection. Specifically, latent variable models, such as principal component analysis (PCA) and partial least squares (PLS) (Ferrer-Riquelme, 2009; Kourti, 2020; Peres & Fogliatto, 2018; Rossouw et al., 2020) are extensively employed in industry to monitor and analyse process data. Fault identification statistics based on latent variable (LV) models, and the covariance structure of the data, are used to detect abnormal behaviour from process data. Once a fault or abnormality is detected, fault diagnosis must be performed to identify the variable or variables that are responsible for the fault. Given accurate fault diagnosis analysis, engineers can investigate and implement corrective actions.

Many different fault diagnosis methods based on LV models have been proposed in the literature (Kuang et al., 2015; Wang et al.,

* Corresponding author at: Unit for Data Science and Computing, North-West University, Potchefstroom, 2531, South Africa.
E-mail addresses: praisehava@gmail.com (P.O. Obanya), roelof.coetzer@nwu.ac.za (R.L.J. Coetzer), carel.olivier@nwu.ac.za (C.P. Olivier), tanja.verster@nwu.ac.za (T. Verster).

https://doi.org/10.1016/j.cie.2024.110064
Received 6 November 2023; Received in revised form 6 March 2024; Accepted 9 March 2024
Available online 11 March 2024

2017). MacGregor and Cinar (2012) stated that LV models reduce data projection to a lower-dimensional space for fault diagnosis. Two statistics that are often used in combination with LV models are the Hotelling's T^2 statistic and Q charts (Ahmed et al., 2012). According to Ahmed et al. (2012), Hotelling's T^2 statistic is a measurement of major variations and it can detect faulty points in new data if the variation in the latent variables exceeds the variation determined by the model or threshold condition, while the Q statistic, also represented as squared prediction error (SPE), evaluates the goodness of fit of the new sample to the model.

Alcala and Qin (2011) presented and analysed the performance of variable contributions to the Hotelling's T^2 statistic, SPE and combined fault detection statistics, following a PCA model developed on a training or in-control data set. Specifically, from Alcala and Qin (2011), three popular variable contribution methods are complete decomposition contributions, partial decomposition contributions and reconstruction-based contributions. The variables with the highest contributions to the decomposition of the fault detection statistic are classified as the ones responsible for the abnormality in the process. Alcala and Qin (2011) found that the contribution methods based on the SPE and combined statistics are more sensitive to faults compared to the T^2 statistic. Yang et al. (2022) proposed the use of a dynamic concurrent partial least squares (DCPLS) monitoring scheme for large-scale processes. The DCPLS model is based on variable importance in the projection. The authors concluded that the DCPLS model is not only useful for the automatic realization of the division of variables, but also for the characterization of the dynamic information of data.

Permutation entropy (PE) is a pattern recognition tool that is very effective in identifying and determining abnormalities in the dynamics of a system. Following the introduction by Bandt and Pompe (2002), PE has been applied successfully across different fields, including biomedicine (Echegoyen et al., 2020; Yang et al., 2018), econophysics (Alves et al., 2020; Zunino et al., 2009), and physical sciences (Maggs & Morales, 2013; Raath et al., 2022). Although PE has been instrumental in several fields, it has not been applied previously for the purpose of variable contribution analysis in multivariate process monitoring. In this paper, we present PE as a new data-driven method for variable contribution analysis to a fault or abnormal behaviour in a process. Note that faults and fault detection are used in a general context and may refer to any type of irregularity or process upset within a commercial process. However, there can potentially be many different kinds of faults in a commercial process, as will be discussed with the Tennessee Eastman process (TEP).

Process data are typically of a time-series nature and PE quantifies the characteristics of a process variable over time for the in-control data set. Once a fault has been detected, either by a process upset or by any of the multivariate fault detection statistics, PE is applied to compare the characteristics of the variables for the faulty state to that of the in-control data set, using an appropriate number of data points. The PE method for variable contribution analysis presented in this paper is independent of fault detection statistics, and is used to identify variables that undergo the greatest change in their expected behaviour over time, immediately after the fault has been introduced. Hence, the main focus of this paper is to show that PE is capable of identifying and detecting deviations in the process variables due to a fault in the process.

There are several advantages of applying PE for variable contribution analysis over other data-driven methods, such as PCA and PLS. One such advantage of PE is its applicability to non-linear time series without any loss of temporal information (Bandt, 2005; Stosic, 2016). In contrast, in the application of PCA and PLS, one must be cautious against non-linear relationships in the data (Hamrouni et al., 2020; Pirouz, 2006). In addition, PE can be applied to any type of process data, regardless of the distribution it follows. Although PLS is also distributional free (Pirouz, 2006), Rhoads and Montgomery (1996) observed that PCA provides the best results when applied to process data that follows a Gaussian (normal) distribution. Furthermore, Susto et al. (2018) stated that feature extraction, which is characteristic of PLS, may be time-consuming. On the other hand, PE is computationally efficient (Bandt, 2005).

Another issue with the use of PCA is the complexity of its application for fault diagnosis, such as difficulty in obtaining the right number of principal components to represent the system (Rahoma, 2021), and difficulty in interpreting the extracted principal components (Xie et al., 2013). PLS shares this complexity in application due to the difficulty in the interpretation of loadings of independent latent variables (Pirouz, 2006). Furthermore, PCA and PLS lack robustness in dealing with outliers (Keboola, 2022; Pirouz, 2006). In contrast, PE only makes use of the order of values in a time series, which makes it suitable for dealing with outliers (Bandt, 2005).

The computational efficiency of PE makes it easy to implement for variable contribution analysis within a commercial process that generally generates large volumes of data. The application of PE for variable contribution analysis is new in the multivariate process monitoring literature, and we use the TEP simulated data set to illustrate and discuss its effectiveness and limitations in detail. Therefore, this paper adds to the existing literature on both PE and variable contribution analysis. It also presents a new application of the PE method and a new approach to variable contribution analysis. To establish the use of PE for variable contribution analysis, a new formula termed PE criterion (PEC_F), is introduced.

The paper is outlined as follows: Section 2 presents an overview of PE. In Section 3, the TEP simulated data set is discussed. Section 4 presents the methodology used, detailing the steps carried out for the variable contribution analysis. Section 5 explains how PE identifies deviations in process variables due to a fault introduction. In Section 6, the variables that contribute to the faults in the TEP are discussed. Section 7 provides a conclusion on the work.

2. Permutation entropy (PE)

PE is represented by the variable H for calculations. H satisfies the condition $0 \leq H \leq 1$ and it measures the order of fluctuations within a time series. When $H = 0$, it implies that the ordinal patterns of the series are completely predictable, while $H = 1$ implies that the ordinal patterns are completely unpredictable. Thus, the greater the value of H, the less predictable the system is. For the identification of dynamic changes within a system, the members of the series are divided into subsets and the order of the values in each subset is determined. Due to the ability of PE to identify deviations from normal behaviour or randomness, it is an effective tool for identifying a change in the behaviour within a time series. There are two important parameters required in the calculation of H, namely the order of PE (size of each subset of the time series), d and the embedding time delay of PE, τ.

For a given time series of length N, an embedding order $d > 2$ and time delay τ, the total number of subsets obtained from N is given by $N - \tau(d-1)$. The choice of d depends on the data size N and should satisfy the condition $N \geq 5d!$ (Amigó et al., 2008; Riedl et al., 2013). The elements of each subset are assigned a rank number from the set $[1, \ldots, d]$. There are $d!$ possible ranking categories for each subset. The frequency of each category relative to the total number of subsets is used to construct a probability vector P. For a vector of probabilities P of dimension $d!$ with element p_j corresponding to the probability of the jth category to occur, H is given by

$$H(P) = S(P)/\ln(d!), \tag{1}$$

where

$$S(P) = -\sum_{\substack{j=1 \\ p_j \neq 0}}^{d!} p_j \ln(p_j). \tag{2}$$

A detailed description of this calculation is provided by Riedl et al. (2013).

Table 1
Measured and manipulated variables of the TEP.

Variable number	Variable name	Variable number	Variable name
xmeas1	A feed (stream 1)	xmeas27	Component E (stream 6)
xmeas2	D feed (stream 2)	xmeas28	Component F (stream 6)
xmeas3	E feed (stream 3)	xmeas29	Component A (stream 9)
xmeas4	A and C feed (stream 4)	xmeas30	Component B (stream 9)
xmeas5	Recycle flow (stream 8)	xmeas31	Component C (stream 9)
xmeas6	Reactor feed rate (stream 6)	xmeas32	Component D (stream 9)
xmeas7	Reactor pressure	xmeas33	Component E (stream 9)
xmeas8	Reactor level	xmeas34	Component F (stream 9)
xmeas9	Reactor temperature	xmeas35	Component G (stream 9)
xmeas10	Purge rate (stream 9)	xmeas36	Component H (stream 9)
xmeas11	Product separator temperature	xmeas37	Component D (stream 11)
xmeas12	Product separator level	xmeas38	Component E (stream 11)
xmeas13	Product separator pressure	xmeas39	Component F (stream 11)
xmeas14	Product separator underflow (stream 10)	xmeas40	Component G (stream 11)
xmeas15	Stripper level	xmeas41	Component H (stream 11)
xmeas16	Stripper pressure	xmv1	D feed flow (stream 2)
xmeas17	Stripper underflow (stream 11)	xmv2	E feed flow (stream 3)
xmeas18	Stripper temperature	xmv3	A feed flow (stream 1)
xmeas19	Stripper steam flow	xmv4	A and C feed flow (stream 4)
xmeas20	Compressor work	xmv5	Compressor recycle valve
xmeas21	Reactor cooling water outlet temperature	xmv6	Purge valve (stream 9)
xmeas22	Separator cooling water outlet temperature	xmv7	Separator pot liquid flow (stream 10)
xmeas23	Component A (stream 6)	xmv8	Stripper liquid product flow (stream 11)
xmeas24	Component B (stream 6)	xmv9	Stripper steam valve
xmeas25	Component C (stream 6)	xmv10	Reactor cooling water flow
xmeas26	Component D (stream 6)	xmv11	Condenser cooling water flow

Table 2
Faults description in the TEP.

Fault number	Description	Type
0	Normal condition without fault	–
1	A/C feed ratio, B composition constant (stream 4)	Step
2	B composition while A/C ratio is constant (stream 4)	Step
3	D feed temperature (stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
6	A feed loss (stream 1)	Step
7	C header pressure loss - reduced availability (stream 4)	Step
8	A, B and C feed composition (stream 4)	Random variation
9	D feed temperature (stream 2)	Random variation
10	C feed temperature (stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	Unknown
17	Unknown	Unknown
18	Unknown	Unknown
19	Unknown	Unknown
20	Unknown	Unknown

3. Tennessee Eastman Process

The Tennessee Eastman Process (TEP) was propounded by Downs and Vogel (1993). The process serves as a gauge to evaluate the performance of industrial process control and condition monitoring techniques. The process consists of five prime unit operations: a reactor, a product condenser, a vapour–liquid separator, a recycle compressor, and a product stripper. The TEP consists of irreversible exothermic reactions yielding two products — G and H from four reactants — A, C, D and E. An inert component B and a byproduct F make a total of 8 components involved in the reaction.

The TEP simulated data, which is available for download at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6C3JR1, consist of 52 variables, including 41 measured variables (xmeas1 to xmeas41) and 11 manipulated variables (xmv1 to xmv11). These 52 variables are listed in Table 1. The process also comprises of two data sets, namely fault-free (Fault 0) data set, which simulates the chemical process having no fault, and faulty data set consisting of the introduction of 20 faults to the chemical process, shown in Table 2. The fault-free data set consists of 500 simulations for each of the 52 variables. Each simulation consists of 500 samples. Hence, there are 250,000 sample entries for each variable. For the faulty data set, each of the 20 faults consists of 500 simulations of 960 samples per simulation for every variable so that each variable has a total of 480,000 sample entries for each of the 20 faults. The fault is introduced from the 161st entry of each simulation, essentially making the first 160 samples fault-free. Note that in the TEP literature, a fault refers to any type of process upset, deviation or equipment failure. From Table 2, the different faults listed are step changes, random fluctuations and a slow drift in the process, as well as a sticking valve. The latter is the only fault that may be classified as an actual equipment failure on the plant. On the other hand, the first three faults are not necessarily caused by equipment or physical faults on the plant. However, in the TEP and variable contribution literature, these types of process deviations or upsets are also referred to as faults (He et al., 2012; Wang et al., 2017), and therefore we use the same language commonly used and accepted

in this domain. Table 2 also includes faults that are unknown and no additional information is available for those.

In this paper, our focus is on variable contribution analysis given a process or performance deviation. Therefore, we consider all faults as process deviations and apply PE for variable contribution analysis of each specific fault.

4. Methodology

In this study, we introduce PE as a new method for variable contribution analysis in process monitoring. To show the efficiency of PE for variable contribution analysis, the TEP data set is used for analysis. There are two sets of data provided: a fault-free data set and a faulty data set, each containing 500 simulations for 52 variables. The fault-free data set is also known as Fault 0, while the faulty data set consists of Faults 1–20.

The fault-free data serves as an in-control data set, and a critical (threshold) value is calculated for H for each of the 52 variables. The H values from a faulty data set are then compared to the in-control critical values to determine the variables that are most affected by the introduction of the fault.

The methodology will now be explained in detail. We use the TEP simulated data as outlined in Section 3. For all calculations of H, we use $d = 3$, which satisfies the condition of $N \geq 5d!$ with $N = 120$. The number of observations was chosen after extensive investigations, which revealed that $N = 120$ yields accurate estimates of H for the variables in the fault-free state. The sample values of the majority of the variables are repeated at every second time stamp. These repeats result in a lower H for the in-control set, and limit the ability of PE to distinguish between faulty and fault-free processes. Thus, to improve the ability of PE to identify changes in the behaviour of the variables, $\tau = 2$ is used in this paper.

From the fault-free data set, H values for each of the 52 variables are obtained. Specifically, for each variable, a total of 381 H values are obtained per simulation. The H values were calculated for the following series of samples for each simulation: 1–120, 2–121, 3–122, 4–123, ..., 379–498, 380–499 and 381–500. This yielded a total of 190,500 (381 for each simulation multiplied by 500 simulations) H values for each variable. The critical value, H_α for each variable was then obtained from the empirical distribution function as the α% percentile from the 190,500 H values.

For each of the faults considered from the faulty data set, a total of 500 H values are obtained for each variable i.e., a H value from each simulation. Specifically, we used the first 120 samples after the introduction of the fault for each variable i.e., samples 161–280, from each simulation for the calculation of H. $N = 120$ is used to calculate H for the faulty data to maintain consistency with the number of samples used for the fault-free data. Since lower values of H indicate increased predictability, H values obtained from the faulty data set are compared to the critical values, H_α, obtained from the fault-free data set, to indicate potential changes in the dynamics of the variables after the fault has been introduced. An empirical analysis gives the percentage of the 500 H values of the variable that falls below its corresponding fault-free critical value. The results are plotted using bar graphs to visualize the contributions of each variable to the faults. The variables with the highest percentage of H values below their corresponding critical values, H_α are the ones that have experienced the greatest change after the fault has been introduced, and consequently contribute the most to the fault.

It is important to note that variables can either be classified as contributing to the fault or be affected by the fault. Therefore, we use contributing and affected interchangeably, since in practice, the analysis of the variable contributions will depend on the type of fault, and often the interest is merely the identification of the effect that a specific fault has on the process variables. Understanding the fault, and the variables affected, are very important information for the engineers in order to bring the process back to normal operation.

The methodology can be summarized as follows:

- Step 1: Let $z^{IC}_{u,j,s}$ be the values of the J number of variables for the IC data set for each of the S simulations, then $u = 1, 2, \ldots, M$ represents the observation number where M is the number of observations, $j = 1, 2, \ldots, J$ represents the variable number where J is the number of variables for the in-control (IC) data set and $s = 1, 2, \ldots, S$ represents the simulation number where S is the number of simulations.
  For calculation of the H values, let $h^{IC}_{k,j,s}, k = 1, 2, \ldots, K, j = 1, 2, \ldots, J, s = 1, 2, \ldots, S$, be the values of H for each variable for the IC or fault-free data set for each of the S simulations. K denotes the number of sequences of length N that can be specified from the time series data of each variable for calculation of the H values.
- Step 2: Obtain the critical value of each variable, H_α as the α% percentile from the empirical distribution function of the $h^{IC}_{k,j,s}$ values. Let $h^{(j)}_\alpha, j = 1, 2, \ldots, 52$ denote the critical values for each variable. Specifically, let $h^{IC}_{(n)}$ be the sorted values from smallest to largest, then $h^{(j)}_\alpha = h^{IC}_{(\alpha n)}$. The critical value, $h^{(j)}_\alpha$ serves as the threshold value in the steps that follow.
- Step 3: For each fault considered, calculate the H values for the faulty data for each variable. Let $z^{F}_{u,j,s}, u = T + 1, \ldots, T + N, j = 1, 2, \ldots, J, s = 1, 2, \ldots, S$, be the values of the J number of variables for the faulty data set for each of the S simulations. T denotes the sample point or time stamp in the time series where the fault was introduced. Then the values of H for each variable for the faulty data set for each of the S simulations, is given by $h^{F}_{j,s}, j = 1, 2, \ldots, J, s = 1, 2, \ldots, S$. Specifically, for each fault considered, $h^{F}_{j,s}$ are calculated for $z^{F}_{161,j,s}$ (the point immediately after the fault was introduced) to $z^{F}_{280,j,s}$, with $T = 160$ and $N = 120$, for each of the 52 variables for all simulations. Therefore, 500 H values are obtained for each variable. The H values obtained from the faulty data set are then compared to the critical values obtained in Step 2 for each variable.
- Step 4: Calculate the percentage of the $S = 500$ H values from the faulty data set that falls below its corresponding fault-free critical value for each variable, to assess variable contribution to the fault.
  The PE Criterion (PEC) for variable contribution to a fault, PEC_F, is given by:

$$PEC_F(j) = \frac{\sum_s I\{h^F_{j,s} \leq h^{(j)}_\alpha\}}{S} \times 100, \quad j = 1, 2, \ldots, J \tag{3}$$

  where I denotes the indicator function. Note the $PEC_F(j)$ is calculated using the critical values $h^{(j)}_\alpha$ obtained from the fault-free data set. It provides the percentage of $h^F_{j,s}$ that is smaller than or equal to the critical value $h^{(j)}_\alpha$, at a specified α. Therefore, it yields a value in the interval [0, 100]. Higher $PEC_F(j)$ indicates a more significant contribution of variable j to the fault.

5. Dynamics of the raw data values and H as transition from fault-free to faulty occurs

In this section, we present the critical values, $h^{(j)}_\alpha$ (using $\alpha = 0.01$) for each of the 52 variables. We also discuss how H and the raw data values react to a fault introduction. In addition, we compare the empirical cumulative distribution function (ECDF) of the H values for the fault-free state to the faulty state across the 500 simulations.

The critical values, $h^{(j)}_\alpha$ for all the variables obtained from the fault-free data set are shown in Table 3. The critical values range from 0.4354 $\left(h^{(38)}_\alpha\right)$ for Component E of stream 11 (xmeas38) to 0.9679 $\left(h^{(14)}_\alpha\right)$ for the product separator underflow (xmeas14). Considering that higher values of H indicate normal or random behaviour, it is expected that the critical values should be close to one for all the variables during in-control operation or a fault-free state. However, Table 3 shows that variables xmeas37 to xmeas41 have low critical

Table 3
Critical values of the variables.

Variable number	Variable name	Critical value, $h_\alpha^{(j)}$	Variable number	Variable name	Critical value, $h_\alpha^{(j)}$
xmeas1	A feed (stream 1)	0.9361	xmeas27	Component E (stream 6)	0.9327
xmeas2	D feed (stream 2)	0.9620	xmeas28	Component F (stream 6)	0.9350
xmeas3	E feed (stream 3)	0.9661	xmeas29	Component A (stream 9)	0.9333
xmeas4	A and C feed (stream 4)	0.9657	xmeas30	Component B (stream 9)	0.9347
xmeas5	Recycle flow (stream 8)	0.9675	xmeas31	Component C (stream 9)	0.9294
xmeas6	Reactor feed rate (stream 6)	0.9659	xmeas32	Component D (stream 9)	0.9347
xmeas7	Reactor pressure	0.9060	xmeas33	Component E (stream 9)	0.9348
xmeas8	Reactor level	0.9673	xmeas34	Component F (stream 9)	0.9300
xmeas9	Reactor temperature	0.9058	xmeas35	Component G (stream 9)	0.9287
xmeas10	Purge rate (stream 9)	0.9476	xmeas36	Component H (stream 9)	0.9314
xmeas11	Product separator temperature	0.9665	xmeas37	Component D (stream 11)	0.4432
xmeas12	Product separator level	0.9676	xmeas38	Component E (stream 11)	0.4354
xmeas13	Product separator pressure	0.9196	xmeas39	Component F (stream 11)	0.4432
xmeas14	Product separator underflow (stream 10)	0.9679	xmeas40	Component G (stream 11)	0.4432
xmeas15	Stripper level	0.9667	xmeas41	Component H (stream 11)	0.4432
xmeas16	Stripper pressure	0.9199	xmv1	D feed flow (stream 2)	0.9642
xmeas17	Stripper underflow (stream 11)	0.9663	xmv2	E feed flow (stream 3)	0.9627
xmeas18	Stripper temperature	0.8083	xmv3	A feed flow (stream 1)	0.9392
xmeas19	Stripper steam flow	0.8174	xmv4	A and C feed flow (stream 4)	0.9666
xmeas20	Compressor work	0.8130	xmv5	Compressor recycle valve	0.9615
xmeas21	Reactor cooling water outlet temperature	0.9410	xmv6	Purge valve (stream 9)	0.9463
xmeas22	Separator cooling water outlet temperature	0.9655	xmv7	Separator pot liquid flow (stream 10)	0.9676
xmeas23	Component A (stream 6)	0.9300	xmv8	Stripper liquid product flow (stream 11)	0.9666
xmeas24	Component B (stream 6)	0.9306	xmv9	Stripper steam valve	0.6100
xmeas25	Component C (stream 6)	0.9326	xmv10	Reactor cooling water flow	0.9590
xmeas26	Component D (stream 6)	0.9288	xmv11	Condenser cooling water flow	0.9663

values. Hence, they do not exhibit regular stochastic behaviour, which indicates predictability during the fault-free state. These variables do not yield a noticeable change in the distribution of H between the fault-free and faulty state. The variables xmeas37 to xmeas41 represent the composition of stream 11, which is the product stream, and might be highly correlated naturally. For these types of process variables, it might be more appropriate to first transform the components to log ratios (Aitchison, 1982), before performing PE. However, this is left for future research.

An investigation into the raw data of all 20 faults, shows that H exhibits consistent changes in certain variables after the introduction of some specific faults. Hence, H is able to distinguish between the contributing and non-contributing variables for these faults. There are three distinct behaviours exhibited by the respective contributing variables of these faults, following the fault introduction. For these three cases, the changes in the raw data of the identified variables are accompanied by significant changes in H. For variables that do not contribute to these specific faults, referred to as Case 4, there is no significant change in the behaviour of the variables, so that H does not exhibit any noticeable reduction after the introduction of the fault. There is an additional case, which involves faults where H was ineffective in identifying the contributing variables, and all 52 variables of these faults show no reduction in H following a fault introduction. Following the introduction of a fault, the five different cases are described as follows:

- Case 1: The values of the contributing variable increase significantly, while H decreases steadily.
- Case 2: The values of the contributing variable decrease significantly, while H decreases steadily.
- Case 3: The values of the contributing variables exhibit a significant increase in variability, while H decreases steadily.
- Case 4: The values exhibit no significant change in the non-contributing variable, while H does not change consistently.
- Case 5: There is no significant change in the values of all the variables after the introduction of the fault, and H does not exhibit any consistent behaviour across different simulations.

To discuss how H and the raw data values react to a fault introduction, we utilize the faulty data set. Specifically, Faults 2, 4, 11 and 14 for Simulations 1, 377 and 500 are considered, which are chosen for illustrative purposes and are considered collectively. For Fault 2, we consider the purge rate (xmeas10) and the stripper steam flow (xmeas19), while for Fault 4, Component E of stream 9 (xmeas33) is considered. For Faults 11 and 14, we consider Component F for stream 6 (xmeas28) and the reactor temperature (xmeas9), respectively. From the faulty data set, the first 280 samples of the simulations of the selected variables are analysed to assess the changes in the values over time, as the samples transition from a fault-free state, represented by the first 160 samples, to a faulty state for the next 120 samples. We also analyse how H changes over time as the fault is introduced. More specifically, the H values for the first 280 samples of each simulation of the faulty data are considered. The H values for $z^F_{k,j,s}, z^F_{k+1,j,s}, \cdots, z^F_{k+N-1,j,s}$, $k = 1, 2, \ldots, 161$ are obtained, with $N = 120$, yielding a total 161 H values. Note that the first 41 H values consist entirely of fault-free samples as they are the H values from samples 1–160. Each of the five cases is illustrated below. Note the legends: Sim 1, Sim 377 and Sim 500, as depicted in Figs. 1(a)–5(a) in the following sections represent Simulations 1, 377 and 500 of the TEP.

5.1. Case 1: The values of the contributing variable increase significantly, while H decreases steadily

An example of a variable that experiences a significant increase after the fault introduction is the purge rate (xmeas10) for Fault 2. Fault 2 represents a step change in the composition of the B feed. Fig. 1(a) shows the dynamics of the raw data values and H, during the transition from a fault-free state to a faulty state. From the top panel of Fig. 1(a), it is observed that the first 160 samples for the illustrative Simulations 1, 377 and 500 vary within a narrow range, from 0.3092 to 0.3697. After the fault introduction, the purge rate increases steadily from 0.3234 to a high of 0.7842. Therefore, the purge rate is clearly affected by the introduction of the fault. From the bottom panel in Fig. 1(a), it is observed that the H values for each simulation starts to decrease immediately after the introduction of Fault 2. Specifically, the first 41 points range from 0.9670 to 0.9983. As the fault is introduced, H declines over time to a low of 0.8101.

To further illustrate the use of H for variable contribution analysis, Fig. 1(b) depicts the ECDF of H for the purge rate (xmeas10), before and after the introduction of Fault 2 for all 500 simulations. From Fig. 1(b), it is clear that H is significantly smaller for the faulty samples

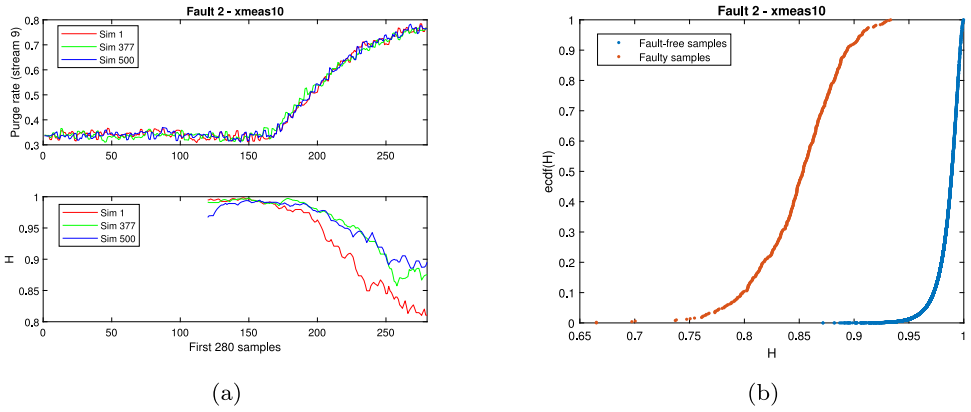

Fig. 1. Illustrative Example for Case 1. (a) The top tile shows the time evolution of the purge rate (xmeas10) for Fault 2, while the bottom tile shows the corresponding behaviour of H. (b) Comparison of the ECDF of H between the fault-free and faulty samples of the purge rate (xmeas10) from Fault 2 for all simulations.

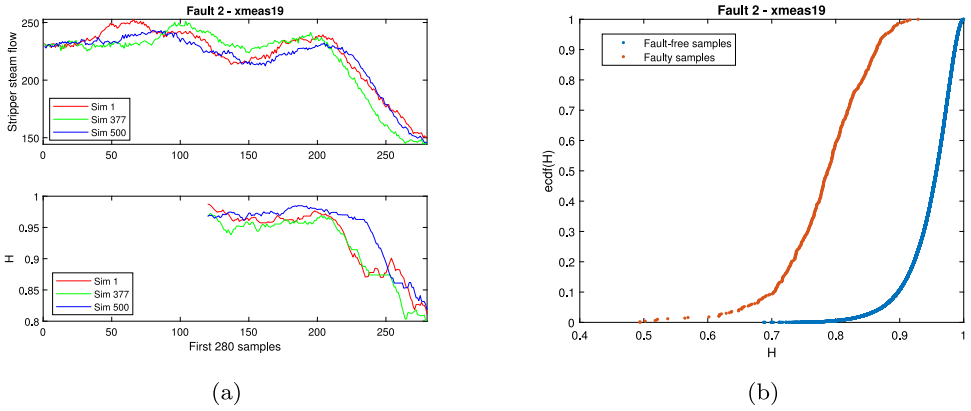

Fig. 2. Illustrative Example for Case 2. (a) The top tile shows the time evolution of the stripper steam flow (xmeas19) for Fault 2, while the bottom tile shows the corresponding behaviour of H. (b) Comparison of the ECDF of H between the fault-free and faulty samples of the stripper steam flow (xmeas19) from Fault 2 for all simulations.

compared to the fault-free samples, and the probability of $H \leq h_{\alpha}^{(10)}$ is equal to one for the faulty samples.

5.2. Case 2: The values of the contributing variable decrease significantly, while H decreases steadily

An example of this case is the stripper steam flow (xmeas19) from Fault 2. The dynamics of the raw data values and H, during the transition from a fault-free state to a faulty state are shown in Fig. 2(a). From Fig. 2(a), it is observed that the first 160 samples from Simulations 1, 377 and 500 vary in quite a narrow range, from 213.24 to 252.8, while the last 120 samples decrease quickly from a high of 241.33 to a low of 144.29. H also changes accordingly, decreasing steadily from a high of about 0.98 at the point where the fault was introduced, to a low of 0.8008. Therefore, H clearly shows the change in the characteristics of the stripper steam flow after the introduction of the fault.

Fig. 2(b) depicts the ECDF of H for the stripper steam flow (xmeas19), before and after the introduction of Fault 2 for all simulations. Fig. 2(b) clearly shows that H is significantly smaller for the faulty samples compared to the fault-free samples, and $\mathrm{P}\left(H < h_{\alpha}^{(19)}\right)$ is approximately 1 for the faulty samples.

5.3. Case 3: The values of the contributing variables exhibit a significant increase in variability, while H decreases steadily

The reactor temperature (xmeas9) from Fault 14 illustrates Case 3. Fault 14 is a sticky reactor cooling water valve, which is expected to influence the reactor temperature. Fig. 3(a) shows the dynamics of the raw data values and H from Simulations 1, 377 and 500 during the transition from a fault-free state to a faulty state, for the reactor temperature. As observed from Fig. 3(a), prior to the introduction of the fault, the raw data values vary in a very narrow range. However, after the fault introduction, there is a significant increase in the size of fluctuations of the reactor temperature. In the bottom panel of Fig. 3(a) it is shown that, the H values decrease consistently from the point where the fault was introduced, from a high of 0.9717 to a low of 0.7238. This clearly demonstrates that H detects the fault in the reactor temperature.

Fig. 3(b) depicts the ECDF of H for the reactor temperature (xmeas9), before and after the introduction of Fault 14 for all simulations. From Fig. 3(b), it is observed that H is significantly smaller for the faulty samples compared to the fault-free samples. The $\mathrm{P}\left(H < h_{\alpha}^{(9)}\right)$ is one for the faulty samples.

5.4. Case 4: The values exhibit no significant change in the non-contributing variable, while H does not change consistently

For variables not contributing to a fault, it is expected that the values of these variables should remain consistent as the process transitions from a fault-free to a faulty state. Accordingly, H should not exhibit any trend or distinctive pattern for these variables, after the introduction of the fault.

As an illustration, Fig. 4(a) depicts the data and H values for the 3 illustrative simulations: 1, 377 and 500, over time for Component F of stream 6 (xmeas28) from Fault 11. From Fig. 4(a), it can be observed that the range of the data over time for the 120 points after the introduction of the fault is similar to the range for the first 160 points for the fault-free data. From the bottom panel in Fig. 4(a), it can

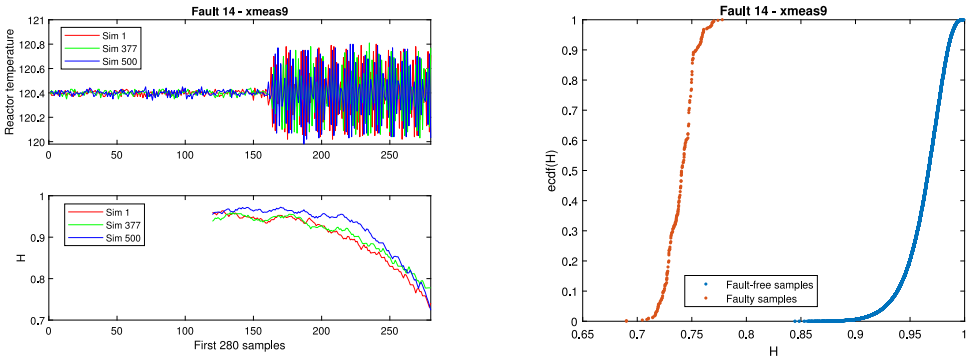

Fig. 3. Illustrative Example for Case 3. (a) The top tile shows the time evolution of the reactor temperature (xmeas9) for Fault 14, while the bottom tile shows the corresponding behaviour of H. (b) Comparison of the ECDF of H between the fault-free and faulty samples of the reactor temperature (xmeas9) from Fault 14 for all simulations.

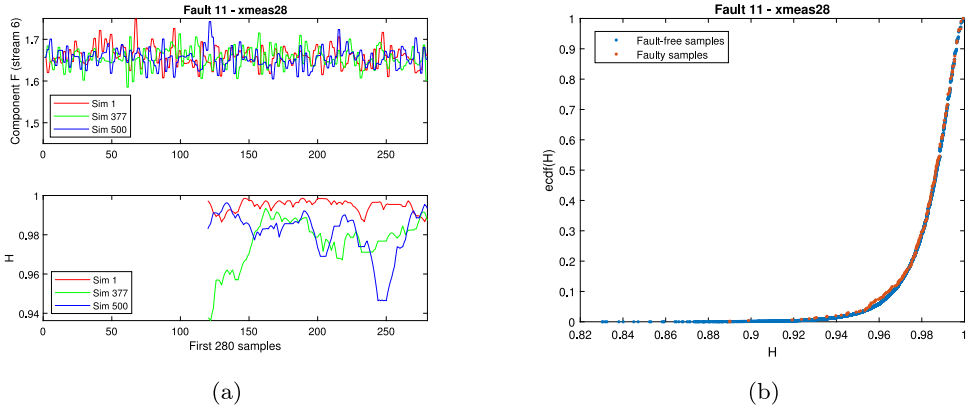

Fig. 4. Illustrative Example for Case 4. (a) The top tile shows the time evolution of Component F of stream 6 (xmeas28) for Fault 11, while the bottom tile shows the corresponding behaviour of H. (b) Comparison of the ECDF of H between the fault-free and faulty samples of Component F of stream 6 (xmeas28) from Fault 11 for all simulations.

be observed that the changes in H are not consistent between the three simulation runs shown. Specifically, there exists great variability in H after Fault 11 was introduced, which indicates that H does not exhibit a clear pattern for variables that do not contribute to a fault.

Fig. 4(b) depicts the ECDF of H for Component F of stream 6 (xmeas28), before and after the introduction of Fault 11 for all simulations. From Fig. 4(b), it is clear that the distribution of H is not different for the faulty samples compared to the fault-free samples, confirming that there is no change in H after the introduction of the fault.

5.5. Case 5: The values exhibit no significant change in the variables and H is ineffective in identifying the contributing variables

This case relates to process faults where H is not capable of identifying the variables responsible for the fault. Specifically, for these types of faults, there is no significant change in the raw values of all the variables after the introduction of the fault. Hence, H exhibits no change between the fault-free and faulty samples. As an illustration, Fig. 5(a) shows the variable plot over time, and the corresponding dynamics of H, for Component E of stream 9 (xmeas33) from Fault 4. From the top panel of Fig. 5(a), it can be observed that the variability is similar between the first 160 fault-free samples and the following 120 faulty samples for Simulations 1, 377 and 500. Consequently, as observed in the bottom panel of Fig. 5(a), H exhibits no consistent change following the introduction of the fault for the three illustrative simulations.

Fig. 5(b) depicts the ECDF of H for Component E of stream 9 (xmeas33), before and after the introduction of Fault 4 for all simulations. From Fig. 5(b), it is clear that the distribution of H is not consistently different for the faulty samples compared to the fault-free samples, confirming that there is little to no change in H after the introduction of the fault.

In conclusion, Cases 1–4 are exhibitions of faults for which H is effective in identifying the contributing variables. For Cases 1–3, where there is evidence of a significant change in the values of the variables that contribute to the fault, PE has the ability to capture variations in the dynamics of the fault-free samples compared to the faulty samples, regardless of how the variations manifest. This ability makes PE an appropriate tool for identifying variables that contribute to faults. The example in Case 4 is typical for variables that do not contribute to the fault under consideration. Case 5 is representative of faults for which H is unable to distinguish between their contributing and non-contributing variables, and therefore all 52 variables exhibit the same behaviour as illustrated.

6. Variable contribution to faults

In this section, we categorize the faults based on the ability of H to determine the variables that contribute to the respective faults for the TEP process. All faults are considered. Subsequently, we compare the results to those from other researchers.

6.1. Efficiency of H for variable contribution analysis

In this section, the PEC_F is used to evaluate the efficiency of PE for variable contribution analysis. The PEC_F obtained for all the variables from all the faults, revealed three categories based on the ability of H

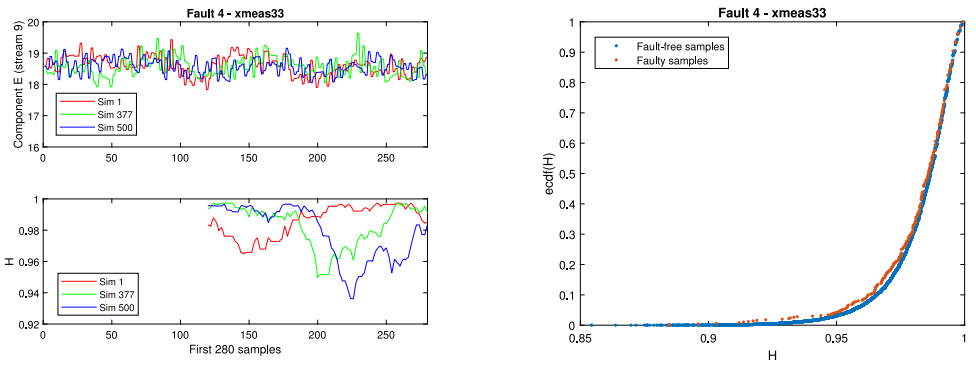

Fig. 5. Illustrative Example for Case 5. (a) The top tile shows the time evolution of Component E of stream 9 (xmeas33) for Fault 4, while the bottom tile shows the corresponding behaviour of H. (b) Comparison of the ECDF of H between the fault-free and faulty samples of Component E of stream 9 (xmeas33) from Fault 4 for all simulations.

Table 4
Categorization of faults based on efficiency of H in determining their variable contributions.

Category	Faults
Category 1	1, **2**, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20
Category 2	**10**, 16
Category 3	3, 4, 9, **15**, 19

to identify variables that contribute to the faults. The three categories are defined as follows:

- Category 1: H strongly distinguishes between the dynamics of the variables during the faulty state and the fault-free state for the variables that contribute to the fault. The variables that are identified as contributing to the faults yield a $PEC_F(j) \geq 90\%$. These variables are referred to as the primary contributors to the faults. The secondary contributors to the fault have a $PEC_F(j)$ between 60% and 90%.
- Category 2: H shows an intermediate level of distinction between the dynamics of the variables during the faulty state and the fault-free state for the variables contributing to the fault. The primary contributors to faults in this category have a $PEC_F(j)$ between 50% and 60%.
- Category 3: H is not effective in distinguishing between the dynamics of the variables during the faulty state and the fault-free state for all the variables. Hence, H is unable to identify the variables that contribute to the faults in this category. All variables (both contributing and non-contributing) have a $PEC_F(j) < 3\%$ for Category 3.

Table 4 lists the categorizations of the faults. Table 4 shows that thirteen of the TEP faults belong to Category 1 and two belong to Category 2. Only five of the TEP faults belong to Category 3. This shows the ability of PE to identify contributing variables to the majority of the different types of faults on the TEP process.

6.2. Examples of the categories of faults

In this subsection, we present an example of each of the categories listed above. Faults 2, 10 and 5 are selected as examples of Categories 1, 2 and 3, respectively.

6.2.1. Example of category 1: Fault 2

Fig. 6 shows the contributions for the top 20 variables based on the $PEC_F(j)$ for Fault 2. From Fig. 6, the purge rate (xmeas10) and the purge valve (xmv6) are identified as the primary contributors to Fault 2. Both variables yield a $PEC_F(j)$ of 100%. The secondary contributor to Fault 2 is identified as the stripper steam flow (xmeas19) with $PEC_F(j) = 68.6\%$. Therefore, for Fault 2, the PEC_F provides a clear indication of relevant variables to consider in order to identify the fault.

6.2.2. Example of category 2: Fault 10

The variable contribution plot to Fault 10 based on $PEC_F(j)$ is shown in Fig. 7. The results clearly show that the stripper temperature (xmeas18) is the primary contributor to Fault 10 with a $PEC_F(j)$ of 56.8%. It is noteworthy that the critical value for the stripper temperature is approximately 0.8083 (from Table 3). Therefore, a lower $PEC_F(j)$ may be expected for this variable. Despite this, H is still capable of identifying the stripper temperature as the primary contributor to Fault 10. It can be observed from Fig. 7 that, although xmeas19 is identified as having the second greatest effect on Fault 10, it has a very low $PEC_F(j)$ of 6.8%. The figure shows that there is a significant difference between the primary contributor to Fault 10 and the other variables in terms of $PEC_F(j)$.

6.2.3. Example of category 3: Fault 15

The $PEC_F(j)$ obtained for all 52 variables after the introduction of Fault 15 were less than 3% with Component G (xmeas40) having the highest $PEC_F(j) = 2.4\%$. The results obtained for the variable contribution to Fault 15 are representative of the results obtained for other faults of Category 3.

6.3. Discussion

In this section, we compare the results of Section 6.2 to results obtained by other researchers. These researchers applied different methods to determine the variables that contribute to the faults in the TEP.

Fault 2 is caused by a step change in the composition of Component B in stream 4, while the A/C ratio is constant. This leads to an increase in the inert Component B, which would directly affect the purge rate (xmeas10) as Component B must exit from purge stream 9. The purge valve (xmv6) of stream 9 is also affected. PE identifies both variables as the primary contributors to Fault 2. In addition, the stripper steam flow (xmeas 19) is identified as a secondary contributor. The variables identified as the primary contributors to Fault 2 are mostly consistent with results obtained by He et al. (2012), Kuang et al. (2015) and Wang et al. (2017). These authors applied a reconstruction-based model, regression techniques and a residual evaluation method, respectively. He et al. (2012) and Wang et al. (2017) only identified the purge rate (xmeas10) as the primary contributor to Fault 2. In addition, Wang et al. (2017) identified the E feed (xmeas3) as a minor contributor to Fault 2. On the other hand, Kuang et al. (2015) identified both the purge rate (xmeas10) and the purge valve (xmv6) as the variables that contribute the most to Fault 2.

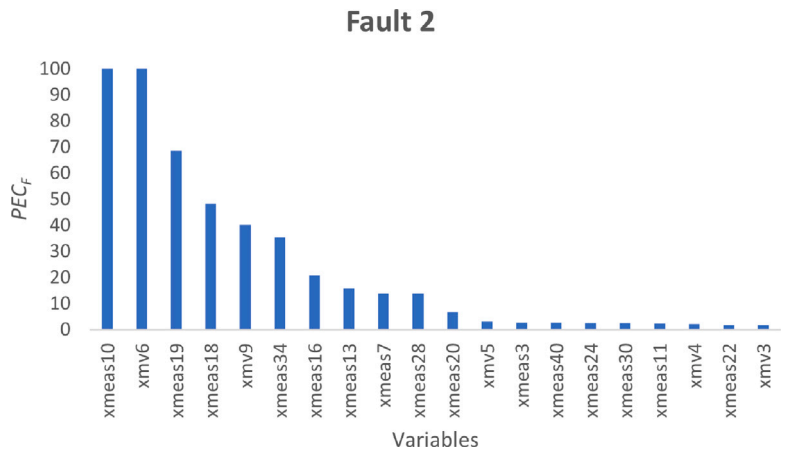

Fig. 6. Variable contribution plot for Fault 2.

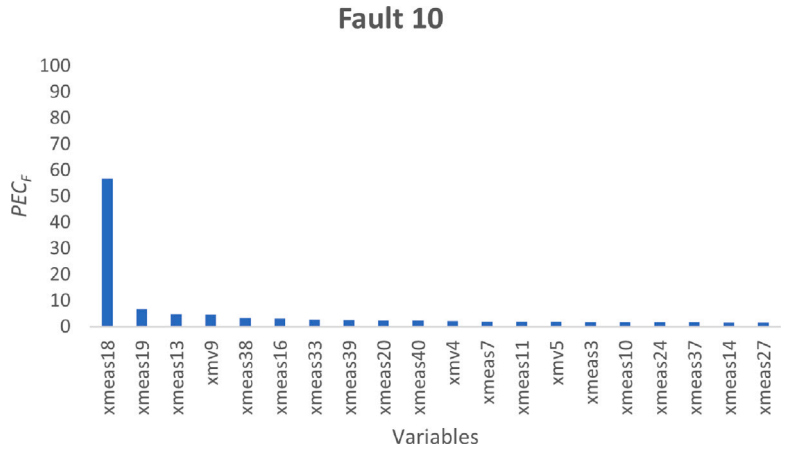

Fig. 7. Variable contribution plot for Fault 10.

Fault 10 is related to a random variation in the C feed temperature. It is known, based on primary knowledge of the TEP, that the stripper temperature (xmeas18) is the root cause of Fault 10 (Zhang et al., 2020). Yang et al. (2022) also noted that the random variation in the C feed temperature causes the stripper temperature (xmeas18) to fluctuate. The stripper temperature (xmeas18) is identified as the primary contributor to Fault 10 based on PEC_F. Results obtained by Li et al. (2014) and Zhang et al. (2020) agree with this result. Li et al. (2014) used a dynamic LV model for analysis, while Zhang et al. (2020) used a PCA-based technique. Both authors identified the stripper temperature (xmeas18) as the variable responsible for Fault 10. The stripper steam flow (xmeas19) was also identified as the secondary contributor to Fault 2 by Li et al. (2014).

Fault 15 is associated with a sticky condenser cooling water valve, and it is acknowledged that it is a challenging fault to detect, with associated variable contributions (Shang et al., 2019; Wang et al., 2016). However, the variable expected to be affected the most by this fault is the condenser cooling water flow (xmv11). This was not identified by PEC_F. This shows the inability of H to accurately determine the variables that contribute to, or is affected by Fault 15. Other faults of Category 3 that has been determined to be difficult to detect are Faults 3 and 9 (Wang et al., 2016).

7. Conclusion

In this paper, we presented permutation entropy (PE) as a new data-driven method for variable contribution analysis in multivariate process monitoring. PE measures the order of fluctuations within a time series, and is independent of fault detection statistics. Unlike other data-driven methods, such as PCA and PLS, PE can be applied to non-linear series, is not affected by outliers, is computationally efficient and easy to interpret. A new formula labelled PE criterion (PEC_F) was specified for variable contribution to a fault. Using the well-known Tennessee Eastman Process (TEP), it was shown that PE can be used effectively to identify and interpret the variables responsible for, or affected by, faults in an industrial process. Therefore, PE can be used for diagnostic analysis and quality control in process and manufacturing industries.

Furthermore, we evaluated the efficiency of PE to identify the primary variables that contribute to the faults simulated for the TEP. The faults were categorized into two groups based on the ability of PE to identify the primary variables contributing to the fault. A third category was also specified for which PE was unsuccessful in identifying the primary variables contributing to the faults. For these faults, there is typically no step change in the values of the variables after the introduction of the fault, or the origin of the fault is unknown. Therefore, further research is required to understand the behaviour of PE, and to improve its effectiveness, in these situations.

This study examines and evaluates the application of PE for the analysis of the contribution of individual variables to faults. Therefore, the effect of the potential correlations between the variables on the PE as a variable contribution analysis technique was not addressed. This effect is referred to as the "smearing effect", and is a well-known problem in variable contribution analysis in latent variable process monitoring (Westerhuis et al., 2000). Recently, there have been many publications that addressed this problem with varying success (e.g., Alcala and Qin 2011, He et al. 2012, Kuang et al. 2015, Shang

et al. 2019). Therefore, since PE can be applied independently of any process monitoring statistic, it is of interest to develop a multivariate PE criterion, which incorporates the correlations between the variables. This will be the focus of our future research. Another aspect to consider in future research is the development of multivariate permutation entropy for fault detection, in combination with the PEC_F criterion for variable contribution analysis in multivariate process monitoring. In addition, an in-depth analysis can be executed to investigate the success rate of PE compared to other existing methods.

CRediT authorship contribution statement

Praise Otito Obanya: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Roelof L.J. Coetzer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Carel Petrus Olivier:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Tanja Verster:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (grant numbers 151037, 145712, 126885).

References

Ahmed, M., Baqqar, M., Gu, F., & Ball, A. (2012). Fault detection and diagnosis using principal component analysis of vibration data from a reciprocating compressor. In *Proceedings of 2012 UKACC international conference on control* (pp. 461–466). IEEE.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *44*(2), 139–160.

Alcala, C. F., & Qin, S. J. (2011). Analysis and generalization of fault diagnosis methods for process monitoring. *Journal of Process Control*, *21*, 322–330.

Alves, L. G., Sigaki, H. Y., Perc, M., & Ribeiro, H. V. (2020). Collective dynamics of stock market efficiency. *Scientific Reports*, *10*(1), 21992.

Amigó, J. M., Zambrano, S., & Sanjuán, M. A. (2008). Combinatorial detection of determinism in noisy time series. *Europhysics Letters*, *83*(6), 60005.

Bandt, C. (2005). Ordinal time series analysis. *Ecological Modelling*, *182*(3–4), 229–238.

Bandt, C., & Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, *88*(17), Article 174102.

De Lázaro, J. M. B., Moreno, A. P., Santiago, O. L., & da Silva Neto, A. J. (2015). Optimizing kernel methods to reduce dimensionality in fault diagnosis of industrial systems. *Computers & Industrial Engineering*, *87*, 140–149.

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers & Chemical Engineering*, *17*(3), 245–255.

Echegoyen, I., López-Sanz, D., Martínez, J. H., Maestú, F., & Buldú, J. M. (2020). Permutation entropy and statistical complexity in mild cognitive impairment and Alzheimer's disease: An analysis based on frequency bands. *Entropy*, *22*(1), 116.

Farahani, S., Brown, N., Loftis, J., Krick, C., Pichl, F., Vaculik, R., & Pilla, S. (2019). Evaluation of in-mold sensors and machine data towards enhancing product quality and process monitoring via Industry 4.0. *International Journal of Advanced Manufacturing Technology*, *105*, 1371–1389.

Ferrer-Riquelme, A. (2009). Statistical control of measures and processes. In *Comprehensive chemometrics* (pp. 97–126). Amsterdam, The Netherlands: Elsevier.

Hamrouni, I., Lahdhiri, H., Abdellafou, K. b., & Taouali, O. (2020). Fault detection of uncertain nonlinear process using reduced interval kernel principal component analysis (RIKPCA). *International Journal of Advanced Manufacturing Technology*, *106*, 4567–4576.

He, B., Yang, X., Chen, T., & Zhang, J. (2012). Reconstruction-based multivariate contribution analysis for fault isolation: A branch and bound approach. *Journal of Process Control*, *22*(7), 1228–1236.

Ji, C., & Sun, W. (2022). A review on data-driven process monitoring methods: Characterization and mining of industrial data. *Processes*, *10*(2), 335.

Keboola (2022). *A guide to principal component analysis (PCA) for machine learning*. Keboola, https://www.keboola.com/blog/pca-machine-learning.

Kourti, T. (2020). Multivariate statistical process control and process control, using latent variables. *Comprehensive Chemometrics*, *4*, 21–54.

Kuang, T., Yan, Z., & Yao, Y. (2015). Multivariate fault isolation via variable selection in discriminant analysis. *Journal of Process Control*, *35*, 30–40.

Li, G., Qin, S. J., & Zhou, D. (2014). A new method of dynamic latent-variable modeling for process monitoring. *IEEE Transactions on Industrial Electronics*, *61*(11), 6438–6445.

MacGregor, J., & Cinar, A. (2012). Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. *Computers & Chemical Engineering*, *47*, 111–120.

Maggs, J., & Morales, G. (2013). Permutation entropy analysis of temperature fluctuations from a basic electron heat transport experiment. *Plasma Physics and Controlled Fusion*, *55*(8), Article 085015.

Peres, F. A. P., & Fogliatto, F. S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers & Industrial Engineering*, *115*, 603–619.

Pirouz, D. (2006). An overview of partial least squares. *SSRN Electronic Journal*.

Raath, J., Olivier, C., & Engelbrecht, N. (2022). A permutation entropy analysis of voyager interplanetary magnetic field observations. *Journal of Geophysical Research, Space Physics*, *127*(6), e2021JA030200.

Rahoma, A. A. (2021). *Fault detection and root cause diagnosis using sparse principal component analysis (SPCA)* (Ph.D. thesis), Memorial University of Newfoundland.

Reis, M. S., & Gins, G. (2017). Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. *Processes*, *5*(3), 35.

Rhoads, T. R., & Montgomery, D. (1996). Process monitoring with principal components and partial least squares. In *Proceedings of the 1996 5th industrial engineering research conference* (pp. 683–686). IIE.

Riedl, M., Müller, A., & Wessel, N. (2013). Practical considerations of permutation entropy: A tutorial review. *The European Physical Journal Special Topics*, *222*(2), 249–262.

Rossouw, R., Coetzer, R., & Le Roux, N. (2020). Variable contribution identification and visualization in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, *196*, 1–12.

Shang, C., Ji, H., Huang, X., Yang, F., & Huang, D. (2019). Generalized group contributions for hierarchical fault diagnosis with group lasso. *Control Engineering in Practice*, *93*, 1–12.

Stosic, D. (2016). *Applications of entropy on financial markets* (Master's thesis), Universidade Federal de Pernambuco.

Susto, G. A., Cenedese, A., & Terzi, M. (2018). Time-series classification methods: Review and applications to power systems data. *Big Data Application in Power Systems*, 179–220.

Wang, J., Ge, W., Zhou, J., Wu, H., & Jin, Q. (2017). Fault isolation based on residual evaluation and contribution analysis. *Journal of the Franklin Institute*, *354*(6), 2591–2612.

Wang, B., Yan, X., & Jiang, Q. (2016). Independent component analysis model utilizing de-mixing information for improved non-Gaussian process monitoring. *Computers & Industrial Engineering*, *94*, 188–200.

Westerhuis, J., S.P., G., & A.K., S. (2000). Generalized contribution plots in multivariate statistical process monitoring.. *Chemometrics and Intelligent Laboratory Systems*, *51*(1), 95–114.

Xie, L., Lin, X., & Zeng, J. (2013). Shrinking principal component analysis for enhanced process monitoring and fault isolation. *Industrial and Engineering Chemistry Research*, *52*(49), 17475–17486.

Yang, J., Wang, J., Sha, J., Dai, H., & Liu, H. (2022). Quality-related monitoring of distributed process systems using dynamic concurrent partial least squares. *Computers & Industrial Engineering*, *164*, Article 107893.

Yang, Y., Zhou, M., Niu, Y., Li, C., Cao, R., Wang, B., Yan, P., Ma, Y., & Xiang, J. (2018). Epileptic seizure prediction based on permutation entropy. *Frontiers in Computational Neuroscience*, *12*, 55.

Zhang, C., Guo, Q., & Li, Y. (2020). Fault detection in the Tennessee Eastman benchmark process using principal component difference based on k-nearest neighbors. *IEEE Access*, *8*, 49999–50009.

Zunino, L., Zanin, M., Tabak, B. M., Pérez, D. G., & Rosso, O. A. (2009). Forbidden patterns, permutation entropy and stock market inefficiency. *Physica A. Statistical Mechanics and its Applications*, *388*(14), 2854–2864.