



Research article

Dynamic monitoring and anomaly tracing of the quality in tobacco strip processing based on improved canonical variable analysis and transfer entropy

Linchao Yang¹, Ying Liu¹, Guanglu Yang^{1,*} and Shi-Tong Peng²

¹ Nanyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd., Nanyang 473007, China

² Key laboratory of intelligent manufacturing technology (Ministry of Education), Shantou University, Shantou 515063, China

* **Correspondence:** Email: ygl9801@sina.com.

Abstract: Multivariate statistical monitoring methods are proven to be effective for the dynamic tobacco strip manufacturing process. However, the traditional methods are not sensitive enough to small faults and the practical tobacco processing monitoring requires further root cause of quality issues. In this regard, this study proposed a unified framework of detection-identification-tracing. This approach developed a dissimilarity canonical variable analysis (CVA), namely, it integrated the dissimilarity analysis concept into CVA, enabling the description of incipient relationship among the process variables and quality variables. We also adopted the reconstruction-based contribution to separate the potential abnormal variable and form the candidate set. The transfer entropy method was used to identify the causal relationship between variables and establish the matrix and topology diagram of causal relationships for root cause diagnosis. We applied this unified framework to the practical operation data of tobacco strip processing from a tobacco factory. The results showed that, compared with traditional contribution plot of anomaly detection, the proposed approach cannot only accurately separate abnormal variables but also locate the position of the root cause. The dissimilarity CVA proposed in this study outperformed traditional CVA in terms of sensitiveness to faults. This method would provide theoretical support for the reliable abnormal detection and diagnosis in the tobacco production process.

Keywords: tobacco processing; canonical variable analysis; transfer entropy; reconstruction-based contribution; anomaly detection

1. Introduction

Current modern light industry has witnessed intelligent transformation towards a smart manufacturing paradigm [1]. The level of quality control and equipment management remains an important indicator for assessing the intelligent manufacturing capabilities [2]. The current industrial big data platform in tobacco workshops makes real-time monitoring of production equipment state and production parameters possible [3]. Line quality monitoring and abnormal diagnosis based on production process data have become key technologies for ensuring high-quality and efficient operation in tobacco workshops [4]. How to fully utilize monitoring data to achieve intelligent root cause also pose a great challenge for tobacco enterprises.

As the core process in tobacco production, tobacco strip processing technology transforms tobacco leaves into qualified tobacco with stable and consistent quality based on the physical and chemical characteristics of the leaf tobacco [5]. It typically includes three major production stages: primary processing, tobacco strip processing and blending and flavoring. Among them, the tobacco strip processing stage is the main process in the entire tobacco process [6]. This stage is a typical dynamic manufacturing process that operates in batch mode, characterized by multiple batches and frequent product changes. Such characteristics of the operation process could scientifically challenge precise monitoring and root cause of quality issue of the tobacco strip processing.

Current research on monitoring and diagnosing the consistency of quality of tobacco strip processing is mostly based on univariate or multivariate statistical process control (MSPC) methods [7,8], using Partial Least Squares (PLS) [9], Principal Component Analysis (PCA) [10] and their extension models, such as Recursive PCA [11], Kernel PCA [12], Multidirectional PLS [13] and Sparse PLS [14], in conjunction with multivariate control charts, such as Hotelling T₂ control chart [15] and Exponentially Weighted Moving Average Control Chart [16] or multivariate generalized likelihood ratio control chart [17], to monitor the anomalies and fluctuations in indicators of tobacco production status. Neural networks [18,19] or contribution plots [20] are then used for anomaly diagnosis. Zhao and Gao [21] developed sparse dissimilarity algorithm to isolate the incipient abnormal variables in cigarette production particularly the cut-made process. Feng et al. [22] proposed a dual attention-based encoder-decoder model as a long short-term memory network to make an online fine-grained quality prediction for the cigarette production process. The primary quality indicators are the moisture content from drier machines. Wang and Zhao [23] combined the nest-loop fisher discriminant analysis and relative change analysis to perform a probabilistic fault diagnosis. They demonstrated the effectiveness of this method via online monitoring and diagnosis of the cigarette cut-made process. Zhao et al. [24] developed an adversarial smoothing regularization for soft sensor and also design a tri-regression framework to improve the generalization performance. They applied this approach to the cigarette manufacturing process to predict the moisture at the outlets of drier machines. Shao et al. [25,26] proposed a bi-dimensional empirical wavelet transform based filtering approach and extended discrete modal decomposition approach to separate irregularities of the manufacturing process.

The above research can effectively improve the control of quality fluctuations within different batches, to a certain extent, controlling batch consistency and stability. Nonetheless, the PLS, PCA and their extension algorithms contain complex singular value decomposition calculations. This calculation would lead to a significant increase in computational complexity with increasing data size, prohibiting real-time monitoring of process quality. Given that modeling with canonical variate analysis (CVA) only requires a one-step singular value decomposition with low computational cost [27],

CVA has been adopted as a typical MSPC approach for nonlinear dynamic process monitoring [28]. However, traditional CVA is not sensitive to incipient small faults and takes longer time to identify faults. This is because the initial faults are usually small enough to be masked by external disturbance, noise or accommodated by the control system. Additionally, traditional contribution plots and neural network methods can preliminarily identify the abnormal-related variables [29]. However, it is impossible to trace the root cause of anomalies and fail to achieve satisfactory monitoring and diagnostic results. The essential reason is that such methods highly depend on the correlation characteristics rather than the causal relationship between variables.

To fill the research gaps above, we drew insight from dissimilarity analysis algorithm from the works by [21] and integrated this concept into traditional CVA. To enhance the sensitiveness of identifying incipient small faults in dynamic processes while ensuring the accuracy of online monitoring, we developed a dissimilarity CVA-based quality monitoring model for the tobacco strip processing. Transfer Entropy (TE) unifies information transfer and signal complexity to describe the causal relationship produced by information flow and is widely used for driving-response relationships in linear or nonlinear time series [30]. Based on the dissimilarity CVA model, this paper established an abnormal root cause diagnosis model based on TE and validated the effectiveness of proposed method using actual operational data from the tobacco production process. The contributions of this paper lie in the following aspects: (i) a unified framework of detection-identification-tracing was proposed for the purpose of identifying the root cause of faults ; (ii) a dissimilarity CVA was developed to avoid the insensitiveness of small faults in traditional CVA; (iii) superior performance of the proposed method was demonstrated via a detailed tobacco strip processing case. This study could arguably offer theoretical support for process state monitoring and intelligent diagnosis in tobacco production to ensure the stable and reliable operation.

The rest of this study is organized as follows: Part 2 described the theoretical framework for the dynamic monitoring and anomaly tracing in tobacco strip processing. Part 3 illustrated the procedure of tobacco strip processing and analyzed the dynamic characteristics. Part 4 provided the model validation and results analysis via a practical case of strip processing in tobacco plant of China. Part 5 discussed and concluded the study.

2. Theoretical framework of anomaly monitoring and root cause of quality issue

The framework for quality detection and diagnosis in tobacco production's dynamic process mostly consists of three parts: offline modeling, online monitoring and identification and root cause diagnosis, as shown in the figure below. Starting with the offline process data of the tobacco strip processing, the dissimilarity CVA algorithm is used to extract high-dimensional data features, establish statistical estimates and control limits. Fault variables are separated adopting the Reconstruction-Based Contribution (RBC) and form a candidate set of quality-related variables for abnormal diagnosis. The TE model was used to quantify the causal relationships between variables in the candidate set and construct a topology diagram of causal relationship for root cause diagnosis.

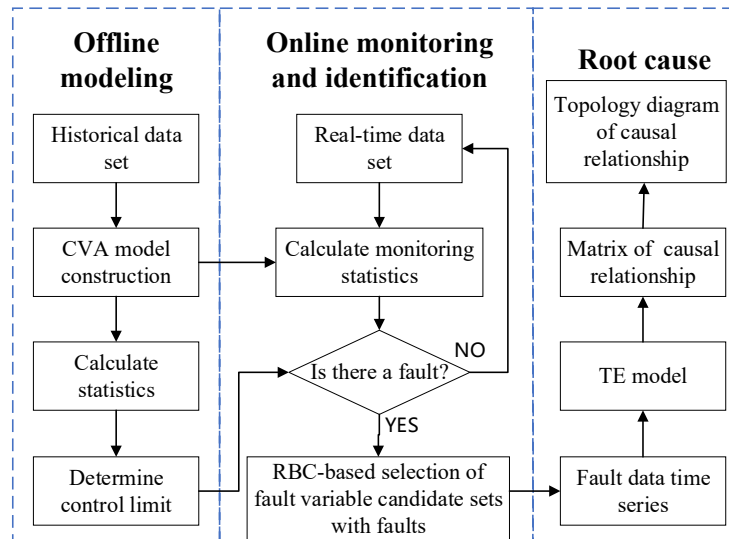


Figure 1. Flowchart of quality monitoring and abnormal diagnosis of tobacco strip processing.

2.1. Revisit the dynamic process modeling based on CVA

CVA is a canonical subspace identification method in multivariate statistical analysis that achieves dimensionality reduction of high-dimensional data by maximizing variable correlations [27]. Assuming that the dynamic characteristics of the system state space model are represented as follows:

$$\begin{cases} \mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{w}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) + \mathbf{v}(t) \end{cases} \quad (1)$$

where $\mathbf{x}(t) \in \mathbf{R}^n$ and $\mathbf{y}(t) \in \mathbf{R}^m$ represent n-dimensional state vector and m-dimensional mass vector, respectively. $\mathbf{u}(t) \in \mathbf{R}^l$ is time series input data of the system, $w(t)$ and $v(t)$ are independent process and measurement white noise respectively; A , B , C and D are the coefficient matrices of the system. In order to decompose time series data, let time t denote the current time, and define the past information vector and present-to-future information vector as:

$$\mathbf{p}(t) = [\mathbf{y}_{t-l}^T, \dots, \mathbf{y}_{t-1}^T, \mathbf{u}_{t-1}^T, \dots, \mathbf{u}_{t-l'}^T]^T \in \mathbf{R}^{l+m} \quad (2)$$

$$\mathbf{f}(t) = [\mathbf{y}_t^T, \mathbf{y}_{t+1}^T, \mathbf{y}_{t+2}^T, \dots, \mathbf{y}_{t+f'}^T]^T \in \mathbf{R}^{f+m} \quad (3)$$

where l' and f' are the number of past and future truncation times, respectively. Based on this, the Hankel matrices of the aforementioned vectors are defined as follows:

$$\mathbf{P} = [\tilde{\mathbf{p}}(t), \dots, \tilde{\mathbf{p}}(t+N-1)] \in \mathbf{R}^{l+m \times N} \quad (4)$$

$$\mathbf{F} = [\tilde{\mathbf{f}}(t), \dots, \tilde{\mathbf{f}}(t+N-1)] \in \mathbf{R}^{f+m \times N} \quad (5)$$

where $\tilde{\mathbf{p}}(t) = \mathbf{p}(t) - \bar{\mathbf{p}}(t)$, $\tilde{\mathbf{f}}(t) = \mathbf{f}(t) - \bar{\mathbf{f}}(t)$, $\bar{\mathbf{p}}(t)$ and $\bar{\mathbf{f}}(t)$ are the sample mean. The covariance matrix and cross-covariance matrix of the two Hankel matrices can be represented as:

$$\sum_{PP} = \frac{1}{N-1} \mathbf{P}\mathbf{P}^T, \sum_{FF} = \frac{1}{N-1} \mathbf{F}\mathbf{F}^T, \sum_{PF} = \frac{1}{N-1} \mathbf{P}\mathbf{F}^T \quad (6)$$

The linear combination of past and present-to-future information vectors $\mathbf{J}^T \tilde{\mathbf{p}}(t)$ and $\mathbf{L}^T \tilde{\mathbf{f}}(t)$, and \mathbf{J}^T and \mathbf{L}^T , can be regarded as the projection directions of vectors, and Pearson correlation is used to express their correlation. The maximum correlation can be achieved by performing singular value decomposition on matrix \mathbf{H} :

$$\mathbf{H} = \sum_{PP}^{-1/2} \sum_{PF} \sum_{FF}^{-1/2} = \hat{\mathbf{J}} \hat{\Sigma} \hat{\mathbf{L}}^T \quad (7)$$

where, the element on the diagonal of $\hat{\Sigma}$ is the characteristic value of \mathbf{H} , corresponding to the degree of correlation, also $\mathbf{J} = \sum_{PP}^{-1/2} \hat{\mathbf{J}}$, $\mathbf{L} = \sum_{FF}^{-1/2} \hat{\mathbf{L}}$.

The CVA method divides the state space into a canonical correlation subspaces and residual subspaces. Since the two statistics often exhibit similar variations in dynamic processes, this paper adopts the statistics and control limits in canonical correlation subspace to monitor the quality of dynamic processes. When the process data follows a normal distribution, the statistical estimates and control limits in canonical correlation subspace can be expressed as [31]:

$$T^2 = \mathbf{p}_t^T \mathbf{J}_q^T \mathbf{J}_q \mathbf{p}_t \quad (8)$$

$$C_s = \frac{q(N^2 - 1)}{N(N - q)} F_\alpha(q, N - q) \quad (9)$$

where q is the selected maximum correlation order, \mathbf{J}_q is the first q row of \mathbf{J} , $F_\alpha(q, N - q)$ represents the F distribution with degrees of freedom, $N - q$ and q , and the confidence level is α . For new sample data i , if $T_i^2 \leq C_s$, it indicates that the system is in a normal state, otherwise, an anomaly occurs.

2.2. Dissimilarity CVA

The traditional CVA is insensitive to small changes, such as sensor drifts, ambient temperature and humidity variations, or process parameter declines. Motivated by the dissimilarity algorithm [21,32] that quantitatively estimates the data distribution difference between fault and normal situations, one can capture small changes by evaluating how well the future canonical values are predictable from the past canonical values. The basic idea in the dissimilarity algorithm was that change of operation state can be reflected via the monitoring distribution of time-series data covering corresponding operation conditions. This suggests that evaluation of differences between the canonical variables (CVs) projected by past and future data can serve as an index for operation condition. The dissimilarity feature can be presented as below:

$$d(t) = L_n \tilde{\mathbf{f}}(t) - \sum_n J_n \tilde{\mathbf{p}}(t) \quad (10)$$

where the t denotes the sample at time t , n is the number of singular values in CVA model, we selected the top n states to reduce L_n and J_n to first n rows, and $\sum_n = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$. The first term in this formula refers to the future projected CVs, while the second term past projected CVs.

Let $Z_p = L_n F$ and $Z_f = J_n P$, the covariance of $d(t)$ can be determined by:

$$\Sigma_{dd} = \frac{1}{N-1} (Z_f - \Sigma_n Z_p) (Z_f - \Sigma_n Z_p)^T = I - \Sigma_n^2 \in R^{n \times n} \quad (11)$$

At normal operation condition, $d(t)$ is zero mean. The dissimilarity index can be formed as Mahalanobis distance of $d(t)$ from zero, namely normalizing the squared of $d(t)$ by Eq (11):

$$D(t) = d(t)^T (I - \Sigma_n^2)^T d(t) \quad (12)$$

Clearly, different from the T^2 index in traditional CVA that only considers the past data, the dissimilarity index $D(t)$ considers information both in the past and future data. Therefore, small changes of operation can be better detected as the departure of current state from the states predicted using past data. As the index is not necessarily satisfy normal distribution, we utilized the kernel density estimation (KDE) to estimate the probability distribution of $D(t)$. Given the significance level of α , the corresponding control limit can be obtained, such that

$$P(x \leq b) = \int_{-\infty}^b \frac{1}{Nh} \sum_{k=1}^M K\left(\frac{x-x_k}{h}\right) dx \quad (13)$$

where x_k represents the samples of x and M is the total number of samples and h denotes the kernel bandwidth. Details regarding KDE can be found in [21]. A fault is detected when the dissimilarity index exceeds the upper control limit (UCL) of $D(t)$.

2.3. Variable selection based on the RBC model

By monitoring production process data using statistical estimates and control limits, once a fault is detected, diagnosis of the fault source is necessary. The contribution plot reflects the degree of influence of variable changes on system stability. By comparing the contribution rates of various variables to the anomaly, the separation of fault variables can be achieved. RBC was first proposed by Alcalá [28]. Its idea is to estimate the nominal normal measurement values of the variables affected by faults, and to identify and estimate the type and impact of the fault, balancing the difference between the fault sample statistics and the minimum value of the reconstructed sample statistics for measuring the contribution of the variable to the fault. For a fault sample, $x=[x_1, x_2, \dots, x_m]^T$, the statistical indicators of the reconstructed value z_i are as follows:

$$index(z_i) = z_i^T \mathbf{M} z_i = \|\mathbf{x} - \xi_i f_i\|_M^2, \quad z_i = \mathbf{x} - \xi_i f_i \quad (14)$$

where f_i is the fault amplitude, ξ_i is the variable direction vector, with the i -th element as 1 and others as 0, M corresponds to z_i in the canonical correlation subspace. Differentiating the above formula with respect to obtain the minimum index (z_i):

$$\frac{d(index(z_i))}{df_i} = -2(\mathbf{x} - \xi_i f_i)^T \mathbf{M} \xi_i = 0 \quad (15)$$

we get:

$$f_i = (\xi_i^T \mathbf{M} \xi_i)^{-1} \xi_i^T \mathbf{M} \mathbf{x} \quad (16)$$

Thus, the expression for the reconstructed contribution value of variable x_i is as follows:

$$C_i^{RBC} = \frac{RBC_i}{\sum_i RBC_i}, \quad RBC_i = \|\xi_i f_i\|_M^2 = \mathbf{x}^T \mathbf{M} \xi_i (\xi_i^T \mathbf{M} \xi_i)^{-1} \xi_i^T \mathbf{M} \mathbf{x} \quad (17)$$

This section takes the reconstructed contribution diagram of the statistics in the canonical correlation subspace as the criterion, namely $M = J_q^T J_q$. If the reconstructed contribution value of variable x_i is greater than the average contribution value, it is selected into the target candidate set of quality-related variables for abnormal diagnosis.

2.4. Construction of quality topology diagram of causal relationship based on TE

The TE model can describe the causal relationship among multiple variables, provided that the time series objects under study satisfy the Markov process [33]. The parameters of tobacco strip processing are only related to the current state, and belong to independent random processes, which meets the requirements of Markov without aftereffect. Let \mathbf{u}_v^p and \mathbf{u}_r^p be the time series input data corresponding to the two elements in the target candidate set of root cause variables; then the TE between variables is given by:

$$T_{u_v|u_r} = \int_D p(\mathbf{u}_v, \mathbf{u}_v^p, \mathbf{u}_r^p) \ln \frac{p(\mathbf{u}_v | \mathbf{u}_v^p, \mathbf{u}_r^p)}{p(\mathbf{u}_v | \mathbf{u}_v^p)} d\mathbf{u} \quad (18)$$

The equation reflects the dynamic information among variables in the target candidate set, where the joint probability $p(\mathbf{u}_v, \mathbf{u}_v^p, \mathbf{u}_r^p)$ represents the dynamic correlation of the variables. Since the joint probability distribution cannot be calculated directly, it can be estimated using kernel density functions. The kernel density estimation function of vector, $[\mathbf{u}_1^p, \mathbf{u}_2^p, \dots, \mathbf{u}_s^p]^T$ is given by:

$$f(u) = \frac{1}{N\Gamma^s \sqrt{\det(\mathbf{S})}} \sum_{i=1}^N K \left[\Gamma^{-2} (\mathbf{u} - \mathbf{u}_i)^T \mathbf{S}^{-1} (\mathbf{u} - \mathbf{u}_i) \right] \quad (19)$$

where \mathbf{S} is the data covariance matrix, $\Gamma = 1.06N^{-1/(4+s)}$ is the bandwidth of the kernel function, which is a \mathbf{K} -Gaussian kernel. To further determine the direction of influence, an influence direction measurement variable is introduced:

$$t_{u_v \rightarrow u_r} = \frac{T_{u_r|u_v} - T_{u_v|u_r}}{\min\{|T_{u_r|u_v}|, |T_{u_v|u_r}|\}} \quad (20)$$

In the above formula, $t_{u_v \rightarrow u_r} > 0$ indicates that \mathbf{u}_v is the dependent variable and \mathbf{u}_r is the causal variable; $t_{u_v \rightarrow u_r} < 0$ means that \mathbf{u}_v is the causal variable and \mathbf{u}_r is the dependent variable; and $t_{u_v \rightarrow u_r} = 0$ indicates that there is no causal relationship between \mathbf{u}_v and \mathbf{u}_r . Based on the above causal relationship judgments, the matrix of causal relationship can be further synthesized to build the topology diagram of causal relationship and diagnose the root cause of abnormalities.

$$\mathbf{t} = \begin{bmatrix} 0 & t_{u_1 \rightarrow u_2} & \cdots & t_{u_1 \rightarrow u_s} \\ -t_{u_1 \rightarrow u_2} & 0 & \cdots & t_{u_2 \rightarrow u_s} \\ \vdots & \vdots & \vdots & \vdots \\ -t_{u_1 \rightarrow u_s} & -t_{u_2 \rightarrow u_s} & \cdots & 0 \end{bmatrix} \quad (21)$$

3. Description of the tobacco strip processing

Tobacco strip processing serves as an important procedure in cigarette production and determines the flavor, style characteristic and final quality of cigarettes [34]. It mostly contains two operation machines, i.e., the HT warming and humidification machine and the SH drier machine. The illustrative plot of machines was shown in Figure 2. The leaf-silk is dampened and inflated by the HT machine to increase the filling rate of leaf-silk and better prepare for the following operation. Mechanical vibration and saturated steam injection in the vibrating body of HT made the leaf-silk fully mix and contact with the steam and absorb a large amount of heat energy. Moisture on the surface of the leaves could be penetrated into the interior and soften itself. The leaves continuously absorb moisture from the surrounding hot steam and then evaporate and vaporize it repeatedly, causing the volume of the woody fibers in the leaves to expand significantly. The SH machine dries the moist leaf-silk rapidly in a barrel heated with conduction-convection of saturated steam. Leaf-silk is fed into the pre-chamber using a vibrating conveyor, and the heat exchanger plate mounted on the inner wall of the cylinder dries, lifts and scatters the leaf-silk. This process increased the curl and elasticity of leaf-silk with the required moisture content and improved the filling capability. Another merit of this process is to mix well the leaf-silk for even distribution of formula ingredients and additives, and thus ensure the basic stability of the composition and flavor of the same batch of cigarettes.

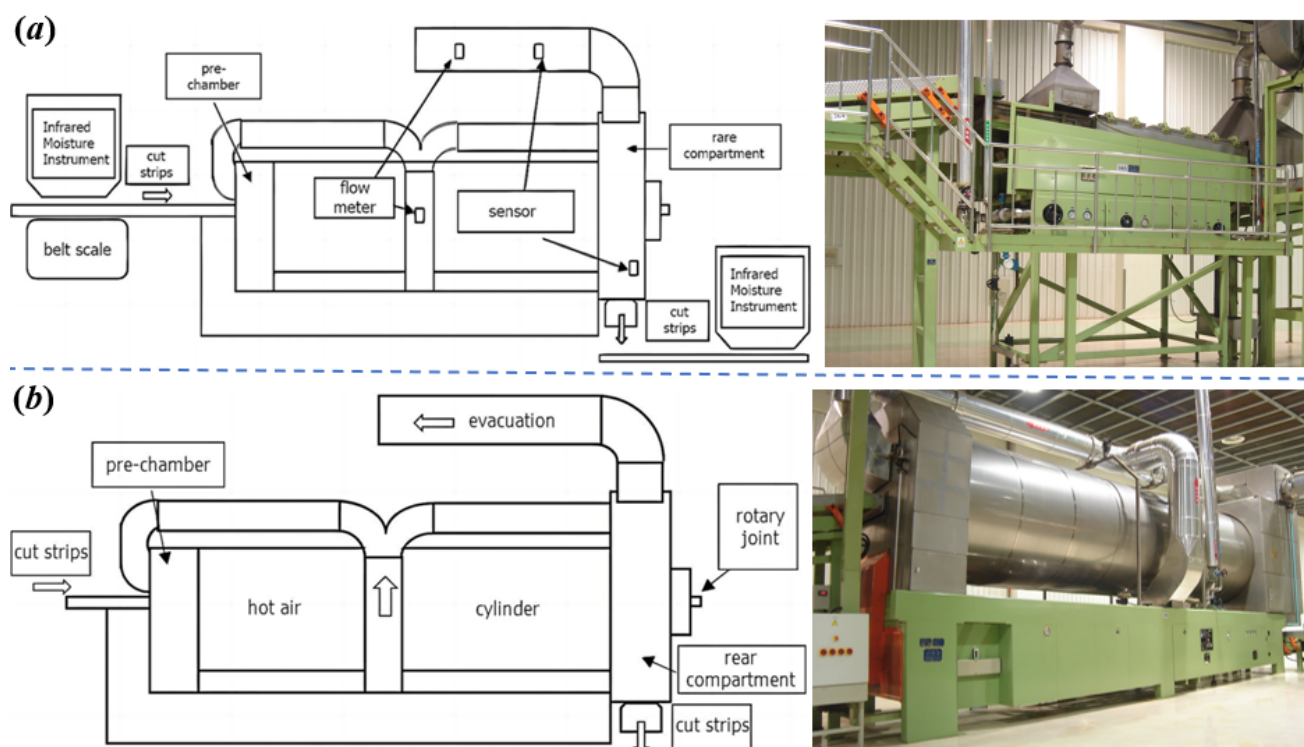


Figure 2. Major machines in tobacco strip processing: (a) HT machine and (b) SH machine.

The tobacco production state data collected by the intelligent monitoring system contains rich quality information. The fluctuation of the quality of the tobacco can be characterized by the dynamic physical data of temperature, humidity and stress. The fluctuation performance of complex and diverse environmental factors, such as high humidity, high temperature and variable pressure at different stages, largely determines the final quality of the tobacco. Due to the numerous quality-related factors and

complex coupling relationships, even a slight anomaly will cause the system to gradually deviate from the initial designed operation state over time [35]. Figure 3 shows the quality fluctuation factors of the tobacco strip processing.

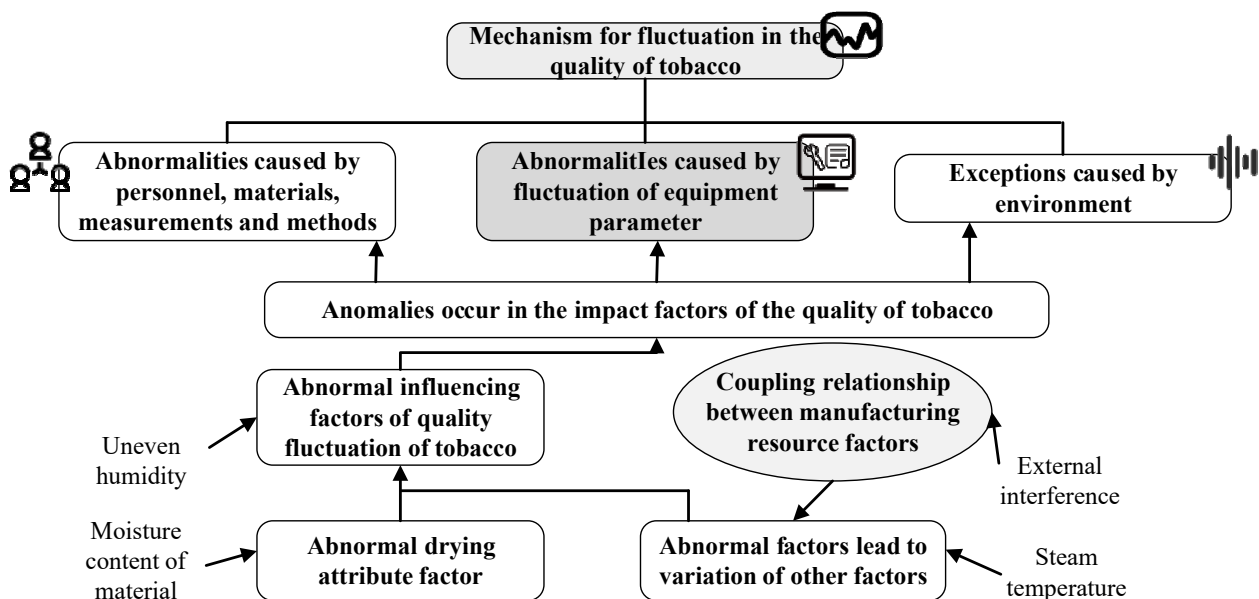


Figure 3. Analysis of fluctuation characteristics of quality of tobacco production.

Table 1. Process variables of the tobacco strip processing.

No.	Variable description	Unit
1	Material flow setting value	kg/h
2	Rotational speed of the cylinder	r/min
3	Opening of the moisture exhaust valve	%
4	Steam flow rate	m ³ /h
5	Material moisture content at HT inlet	%
6	Cylinder wall temperature in zone I	°C
7	Cylinder wall temperature in zone II	°C
8	Steam flow rate at HT	m ³ /h
9	Negative pressure of the exhaust air	μbar
10	Temperature of material at HT outlet	°C
11	Material moisture content at HT outlet	%
12	Temperature of hot air	°C
13	Flow rate of material in the SH	kg/h
14	Flow rate of material in the HT	kg/h
15	Temperature of material at the outlet	°C
16	Steam pressure of the HT	bar

This paper focused on the tobacco production process of Golden Leaf (Dihao Brand) at the Nanyang Tobacco Factory of Henan China Tobacco Industry Co., Ltd. The core equipment on the production line includes SH thin-plate drying equipment and HT temperature-and-humidity-control

equipment, which integrated with various intelligent instruments and sensors, such as temperature and humidity sensors, flow meters, etc. Real-time data was recorded and transmitted to the control terminal, providing rich production process data. The process variables monitored by the model were shown in Table 1, and the process data were obtained from the MES system of the tobacco factory, which can comprehensively describe the status of the system equipment and production. Although some variables are set to their rated values, such as the hot air moisture content of 5500 L/h and the negative pressure of 600 μ bar for dehumidification, the control module kept these variables in a dynamic state. Since the quality characteristics consider both the stability of quality within batches and the consistency of quality among batches, the system needs to be continuously monitored at high frequency over a long period of time. The initial data collection interval was set to 0.5 minutes, and 240 normal operating data samples were continuously collected for 2 hours to establish the model.

The moisture content of the exported material was adopted as an important quality performance indicator. The process standard for the Golden Leaf (Dihao Brand) tobacco product required a moisture content of 12.5–13.5%. The corresponding quality abnormality was that the tobacco at the export stage under the limit of a reasonable moisture content. This article temporarily took this indicator as the quality variable to verify the effectiveness of the proposed quality monitoring-diagnostic model.

4. Model verification and result analysis

4.1. Model parameter setting

In modeling the dynamic process of tobacco strip processing, the quality vector had a dimension of one, and the dimension of the input time series data was 16, with both variables l' and f' set to 3. The confidence level of the F distribution was 95%, and the relevant order q of the projection direction matrix \mathbf{J} was 8. Based on historical operational data, quality monitoring of dynamic process was calculated. To verify the effectiveness of the model, an additional hour of normal operation data, consisting of 120 samples, was collected as test data under the product brand of Golden Leaf (Dihao Brand).

4.2. Online monitoring and abnormal identification

Based on the offline monitoring data samples of two hours of normal operation, statistics and control limits in canonical correlation subspace were constructed, and the projection direction matrix \mathbf{J} was calculated. The process monitoring statistical result was shown in Figure 4(a). The T^2 value has been consistently maintained within the control limit and has a certain distance from the upper and lower control limits, indicating that the system has good process capability. To further test the monitoring ability of the model for system variations, two types of faults were designed, as shown in Table 2. Those faults were brought into the system at the 20th sample time (half minutes). It should be noted that the criterion for determining whether a process is out of control is to exceed the control limit five consecutive times, rather than just one or two occurrences.

Table 2. Fault scenarios in tobacco strip processing.

ID	Variables	Description	Value of δ	Type
F1	Steam flow rate	$R_i = R_{i,0} + \delta t$	0.05	Additive
F2	Material moisture content at HT inlet	$M_i = M_{i,0} + \delta t$	0.005	Additive

*Note: δ shows the rate of fault progress, subscript 0 denotes nominal value and unit of t is half minutes.

In terms of the F1, the results of the T^2 process monitoring are shown in Figure 4(a). The control limits of T^2 and D in Figure 4 were 11.8 and 12.3, respectively. The control chart issued an alarm at the 96th sampling time, accurately detecting the system anomaly after a delay of 38 minutes. While the anomaly can be detected at the 37th sample time, suggesting 8.5 minutes delay when using dissimilarity CVA. The comparison in Figure 4 indicated significant superior performance of dissimilarity CVA in detecting small changes.

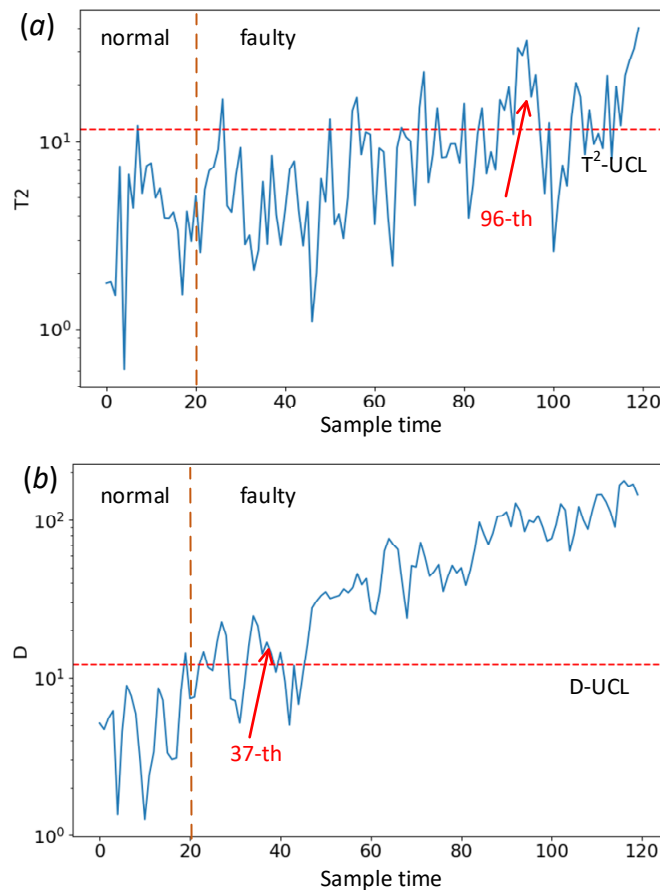


Figure 4. Process monitoring for Fault F1 using: (a) CVA; (b) dissimilarity CVA.

These two methods could effectively monitor abnormal operating conditions in the tobacco strip process. Performance comparison regarding fault F2 was presented in Figure 5. The control limits of T^2 and D in Figure 5 were 18.6 and 13.4, respectively. The dissimilarity CVA detected the anomaly 5 sample time, i.e., 2.5 minutes earlier than the traditional CVA. Even though the dissimilarity CVA had slight better performance, the D index showed a significant increasing trend. The T^2 index in traditional CVA, however, presented a sharp fluctuation along with time. Such a feature could lead to false alarm or correct rejection under specific criteria for identifying anomalies.

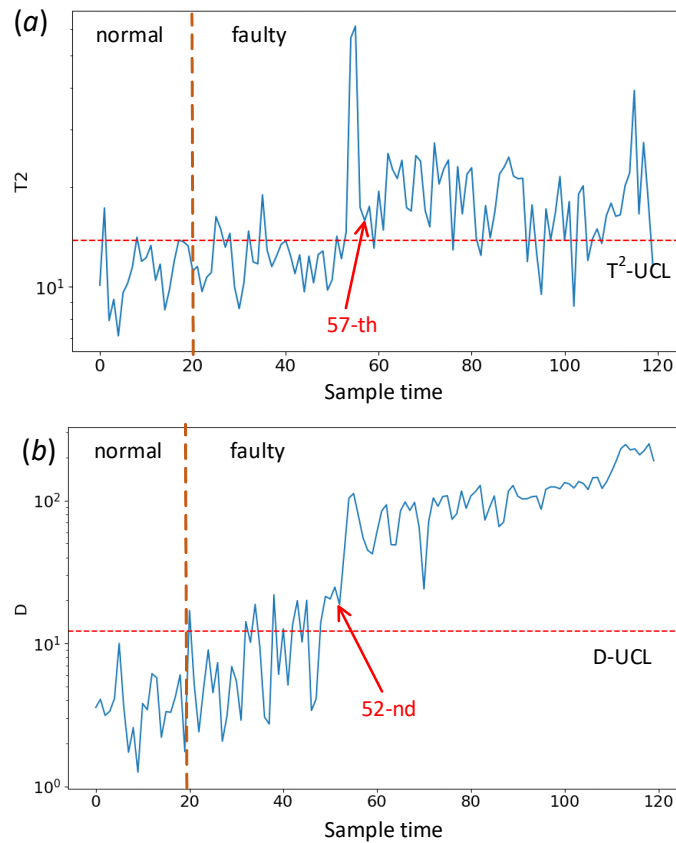


Figure 5. Process monitoring for Fault F2 using: (a) CVA; (b) dissimilarity CVA.

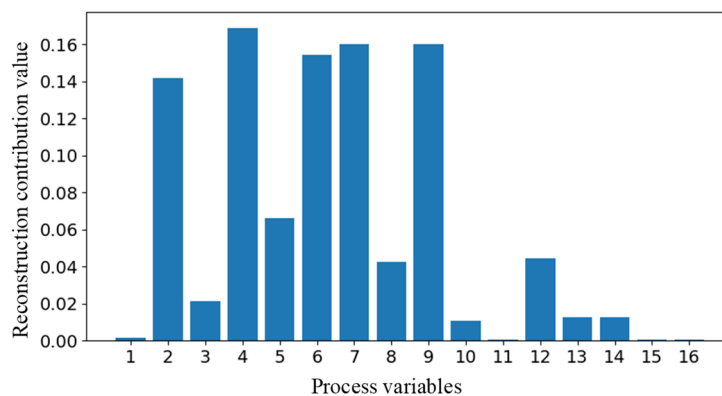


Figure 6. Identification of variable of fault F1 based on RBC.

Figure 6 presented the identification results of fault F1 based on RBC, reflecting the contribution rate of process variables to the changes in statistical estimates. It can be seen that variable 4, the rate of steam flow, has the highest reconstruction contribution rate among the variables in Table 1. Due to the propagation effect of abnormal influences, variables 2, 6, 7 and 9, the rotational speeds of the cylinder, cylinder wall temperature of zone I and zone II, and negative pressure of the exhaust air will also affect the quality. The monitoring results show that the adopted RBC method can effectively separate the relevant fault variables and display their values of contribution rate in real time,

demonstrating good abnormal identification capability. To more precisely identify the root cause of the anomalies, the TE method was used to further quantify the causal relationships between these five variables, as shown in Table 3, by selecting the elements of target candidate set from the variables with high rates of reconstructed contribution in Figure 6. The row numbers in Table 3 correspond to the dependent variables, and the column numbers correspond to the causal variables. “-” indicates that the value of causal relationship between variables is negative or reversed.

Table 3. Matrix of causal relationship of target candidate set of Fault F1.

Variable number	Variable	2	4	6	7	9
2	Rotational speed of the cylinder	0	-	0.426	0.418	-
4	Steam flow rate	0.146	0	0.378	0.353	0.179
6	Temperature of cylinder wall in zone 1	-	-	0	-	0.514
7	Temperature of cylinder wall in zone II	-	-	0.062	0	0.532
9	Negative pressure of the exhaust air	0.211	-	-	-	0

For fault F2, Figure 7 illustrated the fault identification results based on RBC, reflecting the contribution of each variable to the abnormal moisture content of the tobacco strip at the outlet. The results indicate that the operating parameters of the HT humidification and temperature-raising equipment, such as the material moisture content at inlet, rate of working steam flow and material moisture content, have higher contribution rates. To clarify the correlation between the key variables, Table 4 presents the matrix of causal relationship between the five major influential variables, where the material moisture content at inlet is the dependent variable of the other variables. Based on the matrix of causal relationship in Tables 3 and 4, a topology diagram of causal relationship of the variables is constructed, as shown in Figure 8. Variables 4 and 5 can be accurately identified as the root cause variables that triggered the abnormal moisture content of the tobacco strip. The proposed method based on canonical variable analysis and transfer entropy for dynamic monitoring and root cause of quality issue can be used for quality control of the tobacco production process, where the abnormal fluctuations in the moisture content of the tobacco strip at outlet can be analyzed using the statistical quantities and control limits of the canonical correlated subspace, and the root cause of abnormal variables can be traced through a quality causality topology model based on transfer entropy.

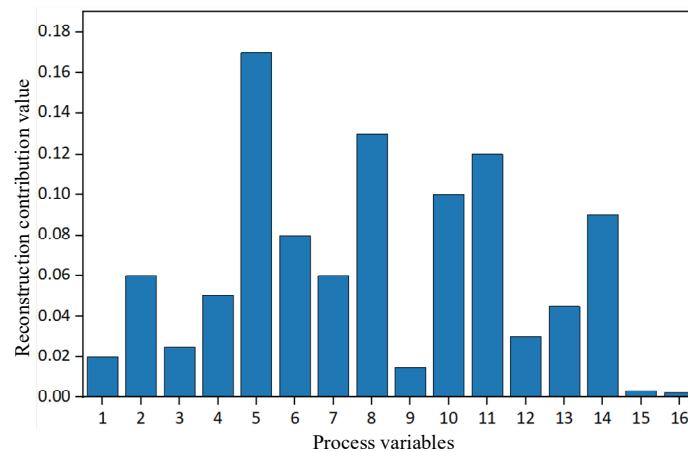
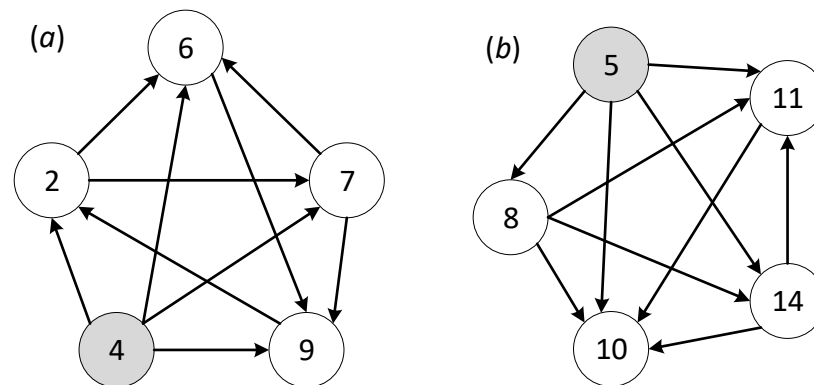


Figure 7. Identification of fault F2 variables based on RBC.

Table 4. Matrix of causal relationship of target candidate set of fault F2.

Variable number	Variable	5	8	10	11	14
5	Inlet material moisture content of the HT	0	0.318	0.139	0.624	0.141
8	Working steam flow rate of the HT	-	0	0.124	0.298	0.527
10	Material temperature at outlet	-	-	0	-	-
11	Material moisture content at outlet	-	-	0.031	0	-
14	Material flow rate	-	-	0.097	0.146	0

**Figure 8.** TE based anomaly root cause diagnosis chart: (a) Fault F1 diagnosis; (b) Fault F2 diagnosis.

5. Conclusions

This paper describes the dynamic process of tobacco production using a state space model and developed a dissimilarity CVA method to fully capture the relationship between process variables and quality variables. The proposed dissimilarity CVA method presented more sensitiveness to slow change faults in the case of tobacco strip processing. As the quality of the tobacco is influenced by multiple variables that exhibit complex correlations, when quality issues arise, the process variables responsible for the anomalies will gradually affect other process variables over time. While fault tree methods can identify potential variables responsible for the anomalies, it is difficult to extract the underlying root causes. This paper adopts reconstruction-based contribution and constructs a target candidate set for quality-related anomalies. The TE method is introduced to mine the correlation information of process variables from historical data, and based on the matrix of causal relationship among the target variables, a topology diagram of causal relationship of process variables is constructed to achieve root cause diagnosis.

The actual operational data of the tobacco production equipment in normal operation are divided into two parts, one for constructing a quality monitoring model and the other for verifying the accuracy of the model. The results of the analysis demonstrate that compared with traditional contribution plot of anomaly detection, the proposed method not only accurately identifies quality issues caused by variable anomalies such as rate of working steam flow and temperature of drum wall, but also determines that the root cause is the rate of working steam flow or material moisture content at HT inlet and other variables, thus improving the reliability of anomaly monitoring and diagnosis.

The method proposed in this paper can be extended to other sections of tobacco production, such

as sections of primary processing or blending and flavoring. It has certain application value for the precise monitoring and tracing of anomalies in the tobacco production and provides a reference for on-site equipment maintenance.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We appreciate the financial support from Special project of Guangdong science and technology innovation strategy (grant No. STKJ2021177 and STKJ202209065) and Technical project of China Tobacco Henan Industrial Co., Ltd. (grant No. RH202002).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. J. Yi, C. Lu, G. Li, A literature review on latest developments of Harmony Search and its applications to intelligent manufacturing, *Math. Biosci. Eng.*, **16** (2019), 2086–2117. <https://doi.org/10.3934/mbe.2019102>
2. J. A. C. Bokhorst, W. Knol, J. Slomp, T. Bortolotti, Assessing to what extent smart manufacturing builds on lean principles, *Int. J. Prod. Econ.*, **253** (2022), 108599. <https://doi.org/10.1016/j.ijpe.2022.108599>
3. X. Shen, Y. Zhang, Y. Tang, Y. Qin, N. Liu, Z. Yi, A study on the impact of digital tobacco logistics on tobacco supply chain performance: taking the tobacco industry in Guangxi as an example, *Ind. Manag. Data Syst.*, **122** (2022), 1416–1452. <https://doi.org/10.1108/IMDS-05-2021-0270>
4. G. W. Vogl, B. A. Weiss, M. Helu, A review of diagnostic and prognostic capabilities and best practices for manufacturing, *J. Intell. Manuf.*, **30** (2019), 79–95. <https://doi.org/10.1007/s10845-016-1228-8>
5. Q. Wang, X. Li, Z. Zhang, B. Tang, Carbon emissions reduction in tobacco primary processing line: A case study in China, *J. Clean. Prod.*, **175** (2018), 18–28. <https://doi.org/10.1016/j.jclepro.2017.11.055>
6. M. Zhu, K. Wu, Y. Zhou, Z. Wang, J. Qiao, Y. Wang, et al., Prediction of cooling moisture content after cut tobacco drying process based on a particle swarm optimization-extreme learning machine algorithm, *Math. Biosci. Eng.*, **18** (2021), 2496–2507. <https://doi.org/10.3934/mbe.2021127>
7. C. Zou, P. Qiu, Multivariate statistical process control using LASSO, *J. Am. Stat. Assoc.*, **104** (2009), 1586–1596. <https://doi.org/10.1198/jasa.2009.tm08128>
8. J. Oakland, R. Oakland, *Statistical Process Control (7th Edition)*, Routledge, New York, 2019. <https://doi.org/10.4324/9781315160511>

9. K. H. Liland, U. G. Indahl, J. Skogholt, P. Mishra, The canonical partial least squares approach to analysing multiway datasets—N-CPLS, *J. Chemom.*, **36** (2022), 1–14. <https://doi.org/10.1002/cem.3432>
10. J. Camacho, A. Pérez-Villegas, P. García-Teodoro, G. MacÍá-Fernández, PCA-based multivariate statistical network monitoring for anomaly detection, *Comput. Secur.*, **59** (2016), 118–137. <https://doi.org/10.1016/j.cose.2016.02.008>
11. P. Qiu, *Introduction to Statistical Process Control*, CRC Press, Boca Raton, 2014.
12. J. Qian, Z. Song, Y. Yao, Z. Zhu, X. Zhang, A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes, *Chemom. Intell. Lab. Syst.*, **231** (2022), 104711. <https://doi.org/10.1016/j.chemolab.2022.104711>
13. X. Wang, P. Wang, X. Gao, Y. Qi, On-line quality prediction of batch processes using a new kernel multiway partial least squares method, *Chemom. Intell. Lab. Syst.*, **158** (2016), 138–145. <https://doi.org/10.1016/j.chemolab.2016.06.017>
14. Q. Jiang, X. Yan, H. Yi, F. Gao, Data-driven batch-end quality modeling and monitoring based on optimized sparse partial least squares, *IEEE Trans. Ind. Electron.*, **67** (2020), 4098–4107. <https://doi.org/10.1109/TIE.2019.2922941>
15. W. Zhou, Z. Zheng, W. Xie, A control-chart-based queueing approach for service facility maintenance with energy-delay tradeoff, *Eur. J. Oper. Res.*, **261** (2017), 613–625. <https://doi.org/10.1016/j.ejor.2017.03.026>
16. C. Zou, W. Jiang, F. Tsung, A LASSO-based diagnostic framework for multivariate statistical process control, *Technometrics*, **53** (2011), 297–309. <https://doi.org/10.1198/TECH.2011.10034>
17. C. Zhao, C. F. Lui, S. Du, D. Wang, Y. Shao, An earth mover's distance based multivariate generalized likelihood ratio control chart for effective monitoring of 3D point cloud surface, *Comput. Ind. Eng.*, **175** (2023), 108911. <https://doi.org/https://doi.org/10.1016/j.cie.2022.108911>
18. M. Dixon, Industrial forecasting with exponentially smoothed recurrent neural networks, *Technometrics*, **64** (2022), 114–124. <https://doi.org/10.1080/00401706.2021.1921035>
19. Y. Wang, M. Perry, D. Whitlock, J. W. Sutherland, Detecting anomalies in time series data from a manufacturing system using recurrent neural networks, *J. Manuf. Syst.*, **62** (2022), 823–834. <https://doi.org/10.1016/j.jmsy.2020.12.007>
20. J. A. Westerhuis, S. P. Gurden, A. K. Smilde, Generalized contribution plots in multivariate statistical process monitoring, *Chemom. Intell. Lab. Syst.*, **51** (2000), 95–114. [https://doi.org/10.1016/S0169-7439\(00\)00062-9](https://doi.org/10.1016/S0169-7439(00)00062-9)
21. C. Zhao, F. Gao, A sparse dissimilarity analysis algorithm for incipient fault isolation with no priori fault information, *Control Eng. Pract.*, **65** (2017), 70–82. <https://doi.org/10.1016/j.conengprac.2017.05.005>
22. L. Feng, C. Zhao, Y. Sun, Dual attention-based encoder-decoder: A customized sequence-to-sequence learning for soft sensor development, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2021), 3306–3317. <https://doi.org/10.1109/TNNLS.2020.3015929>
23. Y. Wang, C. Zhao, Probabilistic fault diagnosis method based on the combination of nest-loop fisher discriminant analysis and analysis of relative changes, *Control Eng. Pract.*, **68** (2017), 32–45. <https://doi.org/10.1016/j.conengprac.2017.07.009>
24. L. Feng, C. Zhao, B. Huang, Adversarial smoothing tri-regression for robust semi-supervised industrial soft sensor, *J. Process Control*, **108** (2021), 86–97. <https://doi.org/10.1016/j.jprocont.2021.11.001>

25. Y. Shao, S. Du, H. Tang, An extended bi-dimensional empirical wavelet transform based filtering approach for engineering surface separation using high definition metrology, *Measurement*, **178** (2021), 109259. <https://doi.org/10.1016/j.measurement.2021.109259>
26. Y. Shao, F. Xu, J. Chen, J. Lu, S. Du, Engineering surface topography analysis using an extended discrete modal decomposition, *J. Manuf. Process.*, **90** (2023), 367–390. <https://doi.org/10.1016/j.jmapro.2023.02.005>
27. B. C. Juricek, D. E. Seborg, W. E. Larimore, Fault detection using canonical variate analysis, *Ind. Eng. Chem. Res.*, **43** (2004), 458–474. <https://doi.org/10.1021/ie0301684>
28. P. E. P. Odiowei, Y. Cao, Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations, *IEEE Trans. Ind. Inf.*, **6** (2010), 36–45. <https://doi.org/10.1109/TII.2009.2032654>.
29. M. Fuentes-García, G. Maciá-Fernández, J. Camacho, Evaluation of diagnosis methods in PCA-based multivariate statistical process control, *Chemom. Intell. Lab. Syst.*, **172** (2018), 194–210. <https://doi.org/10.1016/j.chemolab.2017.12.008>
30. P. Duan, F. Yang, T. Chen, S. L. Shah, Direct causality detection via the transfer entropy approach, *IEEE Trans. Control Syst. Technol.*, **21** (2013), 2052–2066. <https://doi.org/10.1109/TCST.2012.2233476>
31. B. Jiang, D. Huang, X. Zhu, Canonical variate analysis-based contributions for fault identification, *J. Process Control*, **26** (2015), 17–25. <https://doi.org/10.1016/j.jprocont.2014.12.001>
32. M. Kano, S. Hasebe, I. Hashimoto, H. Ohno, Process monitoring based on dissimilarity of time series data, *Kagaku Kogaku Ronbunshu*, **25** (1999), 1004–1009. <https://doi.org/10.1252/kakoronbunshu.25.1004>
33. R. Silini, C. Masoller, Fast and effective pseudo transfer entropy for bivariate data-driven causal inference, *Sci. Rep.*, **11** (2021), 1–13. <https://doi.org/10.1038/s41598-021-87818-3>
34. C. Yuan, P. Yuan, J. Li, Y. Dong, P. Li, The study on the relationship between the cut tobacco drier equipment parameters and the tobacco leaf silk quality, *Stat. Appl.*, **4** (2015), 176–186. <https://doi.org/10.12677/SA.2015.43020>
35. J. Zheng, C. Zhao, Online monitoring of performance variations and process dynamic anomalies with performance-relevant full decomposition of slow feature analysis, *J. Process Control*, **80** (2019), 89–102. <https://doi.org/10.1016/j.jprocont.2019.05.004>



AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)