The International Journal of Advanced Manufacturing Technology (2019) 103:153–163
https://doi.org/10.1007/s00170-019-03520-9

ORIGINAL ARTICLE

A new monitoring scheme of an air quality network based on the kernel method

Maroua Said[1] · Khaoula ben Abdellafou[2] · Okba Taouali[3,4] · Mohamed Faouzi Harkat[5]

Received: 16 November 2018 / Accepted: 21 February 2019 / Published online: 16 March 2019

Abstract
Air pollution is classified as one of the most dangerous type on the human health, the environment, and the ecosystem. However, air pollution results in climate change and affects people's health. For a number of years, monitoring the air quality has become a very urgent and necessary topic. Moreover, safety and health have been attracting attention as one of the important topics to evaluate, firstly, the degree of air pollution and predict pollutant concentrations accurately. Then, it is crucial to establish a more scientific air quality monitoring to ensure the quality of life. In this paper, new reduced air quality monitoring is suggested to enhance the Fault Detection (FD) of an air quality monitoring network. Furthermore, a sensor FD procedure based on Reduced Kernel Partial Least Squares (RKPLS) is proposed to monitor an air quality monitoring network. The main contribution of the suggested procedure is to enhance the FD of an air quality monitoring network in terms of computation time and false alarm rate, using just the important latent components, compared to standard Kernel Partial Least Squares (KPLS).

Keywords Air pollution · Air quality · KPLS · Reduced KPLS · SPE · Fault detection

1 Introduction

Concerning the environment, indoor air pollution is consistently ranked among the top highly important risks that can destroy public health, life quality, vegetation, and even monuments. Because of human activities, industrial effluents, and meteorological factors, air is vulnerable to be polluted by pollutants like nitrogen oxides (NO_2 and NO), ozone, and carbon oxides. Thus, air quality process monitoring is becoming increasingly essential and important to protect the public health and the environment [1–3]. Accordingly, the development of robust and accurate air quality monitoring is highly desirable.

Air quality monitoring networks have been reported in the literature [4–6] to make sure that air quality standards and preventive measures reduce the undesirable change effect in many pollutants. As a consequence, using data of air quality networks is crucial to achieve the desired objectives.

Moreover, the validation and monitoring of sensor networks are very important steps. Generally, the research can be divided to evaluate the air quality monitoring network, according to the following techniques: process modeling, sensor Fault Detection (FD), sensor fault isolation, and correction [7]. In this work, the main purpose of this manuscript is to propose a new technique to improve the sensor FD phase. To ensure good public health,

✉ Okba Taouali
otawali@ut.edu.sa

[1] University of Sousse, National Engineering School of Sousse (ENISO), MARS Research Laboratory, LR17ES05, 4011, Hammam Sousse, Tunisia

[2] Department of Computer Science, Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia

[3] Department of Computer Engineering, Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia

[4] University of Monastir, National Engineering School of Monastir, Monastir, Tunisia

[5] Department of Electronics, Faculty of Engineering Annaba, Badji Mokhtar, BP. 12, 23000 Annaba, Algeria

air quality monitoring has been carried out by several techniques. In the literature [8, 9], monitoring approaches can be divided into two significant categories: model-based and data-driven methods.

For process monitoring, model-based approaches always utilize the process model predictions to make decisions concerning the existence or absence of faults [10–12]. However, several data-based monitoring methods have been developed in the literature [47], such as principal component analysis (PCA) [13–15], independent component analysis (ICA) [16], and partial least squares (PLS) [17, 18]. Several data-based monitoring methods, in the literature, have been used to construct models that can be often utilized in the process monitoring step. The data-based approaches to the monitoring and modeling of real applications, especially air quality networks, usually depend on the quality of used data. Especially, the PCA technique tries to extract linear relations among the considered variables and then represents them with orthogonal principal components. The PCA technique is particularly adapted to reveal linear relationships among the plant variables without formulating them explicitly, as indicated in [19, 20].

However, the PLS, as a data-driven method, has shown good performances and has been widely used in modeling, monitoring, and diagnosis in many fields of chemistry, analytical, physical, and clinical chemistry, as well as the air quality system. The PLS method, which can extract relationships between two sets of variables, inputs/outputs, can build a linear learning model with linear latent variables (LVs) [21]. Unlike the PCA, which captures variations in input data with a descending order of variance, the PLS model finds an optimum pair of latent variables in the input data related to the output ones, such that these transformed variables have the largest covariance.

In this context, several extended PLS techniques have been proposed in the literature. The statistical process based on the PLS method has been frequently studied to have good detection results. In [22], the authors suggested the monitoring methods with multiblock PLS models and showed as well the performance of the FD of the PLS technique. The authors of [23] applied the recursive PLS algorithm in order to update the PLS model with the latest process data. Later, another method, called the total PLS model, was developed by Zhou et al. [24] for output-relevant process monitoring.

From another perspective, the Kernel PLS (KPLS) approaches have become one of the simple, popular, elegant, and fast techniques at the level of the development of the soft measurement model for nonlinear systems relative to other nonlinear approaches [25–27]. The KPLS method provides, at the same time, good monitoring performances by finding those LVs that present a nonlinear correlation with the response variables, besides improving model understanding. The main advantage of the KPLS method is that it does not involve any nonlinear optimization [28] utilizing the kernel function or the stabilization problem, hence making it as simple as standard PLS.

For the KPLS method, the number of latent variables selected for KPLS may be larger than that for linear PLS. However, computation time (CT), selected for the KPLS method, may increase during the identification phase according to the number of samples for the storage of the symmetric kernel matrix of a KPLS monitoring model. To achieve the modeling and monitoring objectives, a new Reduced Kernel Partial Least Squares (RKPLS) method is developed to predict the concentrations of various pollutants and to help understand the modification of air quality networks. The main goal of this paper is to use the advantages of the KPLS technique by introducing it as a part of a new proposed method for air quality networks. Then, the suggested RKPLS method is carried out.

In summary, we develop the RKPLS technique that aims to enhance the data validation of air quality monitoring networks. In this study, the main principle is to treat the problem of CT and the storage of variables. The air quality is a real system that presents several variables. Thus, the treatment of these variables takes a long time. The aspect of the suggested method is to select, using a projection of all LVs, just the important variables. These selected variables consist in presenting a faster and even more effective monitoring process.

In the proposed method, we consider only the set of observations that approximate the retained important components to produce a reduced size of the kernel matrix. Nevertheless, for a large and complex system, our proposal is mainly based on a reduced Gram matrix, so the training time decreases rapidly with a reduced number of observations.

To overcome large datasets, a FD method is developed. The faults sensor or abnormal changes in measured air quality must be detected effectively and quickly using the detection index Squared Prediction Error (SPE). In this task, the FD method uses the SPE index and the Exponentially Weighted Moving Average (EWMA) to improve this phase. The FD performances of the developed RKPLS-based SPE technique and RKPLS-based EWMA-SPE technique are illustrated in terms of False Alarm Rate (FAR), Good Detection Rate (GDR), and CT.

The paper is outlined as follows. The concepts of PLS and KPLS are introduced in Section 2. In Section 3, the proposed RKPLS method is described. Next, Section 4 details the selection of the kernel parameter principle using the Tabu search algorithm. In Section 5, the application results on an air quality monitoring process are given to illustrate the performance of the suggested reduced approach. The conclusion is presented in Section 6.

2 Preliminary

2.1 Standard PLS method

First of all, we are interested in the basic PLS method, which is an extension of the proposed RKPLS method. The general principle is to extract, using the input and output data matrices, LVs to build a linear multivariable model.

The input data $X = \left[x_1, \ldots, x_N\right]^T \in \Re^{N\times m}$ contain N samples with m process variables and the output data $Y = \left[y_1, \ldots, y_N\right]^T \in \Re^{N\times J}$ comprise N observations with J quality variables.

In PLS, we project the input/output data in a low-dimensional space characterized by an L number of latent variables [29]. The PLS method decomposes the input and output data as follows:

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \end{cases} \tag{1}$$

where T=$[t_1, t_2...t_l]$ and U=$[u_1, u_2...u_l]$ are the score vectors, P=$[p_1, p_2...p_l]$ and Q=$[q_1, q_2...q_l]$ are the loadings for X and Y, respectively. Thus, the E and F matrices are, respectively, the PLS residuals of the input data X and the output data Y.

2.2 Theory of the KPLS method

2.2.1 Kernel function

Actually, systems have a nonlinear structure. Because of the limitation of the standard PLS for the nonlinear system, several methods have been developed. According to the trend and popular methods, the kernel technique has received a lot of attention [25, 30].

The KPLS method is characterized by the kernel matrix (Gram matrix) which consists in building nonlinear latent components with an approximately linear computational cost.

Therefore, the main idea is to transform the nonlinear data (input and output) in a higher-dimensional space, called the feature space F, as illustrated in Eq. 2. In this case, the KPLS method is formulated in a feature space of traditional PLS to its nonlinear kernel form [31, 48].

$$\Phi : x_i \in \Re^N \rightarrow \Phi(x_i) \in F \tag{2}$$

Furthermore, we cannot determine the nonlinear mapping of each observation from the batch process. A Mercer kernel $k(.,.)$ is proposed to overtake this problem [32]. Equation 3 presents the product of two mapped samples to determine the kernel function:

$$k(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >= \Phi(x_i)\Phi(x_j)^T \tag{3}$$

where $\Phi(x_i) \in \Re^{1\times S}$, $i = 1, ..., N$ and S is the dimension of the feature space.

In the literature, many kernel functions have been commonly defined and used. Table 1 presents the different kernel functions where p, β_0, β_1, and c are determined using the cross-validation technique.

In a high-dimensional space and prior to calculation, the mean centering of the Gram matrix K must be performed, as presented by Eq. 4:

$$K \leftarrow \left(I_n - \frac{1}{n}1_n1_n^T\right) K \left(I_n - \frac{1}{n}1_n1_n^T\right) \tag{4}$$

where 1_n is a vector of ones whose length is N, and I_n is an N-dimensional identity matrix.

2.2.2 KPLS function monitoring

The PLS kernel algorithm was given by Lindgren et al. [33] with a large number of samples. The Gram matrix K $\in \Re^{N\times N}$ [34] can be presented according to Eqs. 3 and 4, as follows:

$$K = \Phi(X)\Phi(X)^T \tag{5}$$

For nonlinear systems, a traditinal KPLS algorithm is given as follows:

Algorithm 1 KPLS algorithm

```
Input: N×M input data matrix X and N×L output data matrix Y
Output: Input score matrices T, output score matrix U
    Step 1: Calculate kernel matrix and then center;
    Step 2: Set i=1, $K_1 = K$, $Y_1 = Y$;
    Step 3: Random initialized $u_i$ equal to any column of $Y_i$;
    Step 4: $t_i = K_i^T u_i$, $t_i = t_i / \parallel t_i \parallel$;
    Step 5: $c_i = Y_i^T t_i$;
    Step 6: $u_i = Y_i c_i$, $c_i = c_i / \parallel c_i \parallel$;
    Step 7: If $t_i$ converges, go to Step 7; else return to Step 3;
    Step 8: Deflate K and Y;
    Step 9: Repeat steps 3 to 6 to extract more latent variables;
    Step 10: Obtain cumulative matrices T and U.
```

According to Algorithm 1, the deflation step is obtained by the rank-one reduction of K and Y [28]. Using a new T score vector, the K and Y matrices are deflated as:

$$K = K - tt^T - Ktt^T + tt^T Ktt^T \tag{6}$$

$$Y = Y - tt^T Y \tag{7}$$

Table 1 Different kernel functions

Polynomial kernel	Sigmoid kernel	Radial basis kernel
$K(X,Y) =< X, Y >^p$	$K(X,Y) = \tanh(\beta_0 < X, Y > + \beta_1)$	$K(X,Y) = \exp\left(-\frac{\|X-Y\|^2}{c}\right)$

where I_n is an N-dimensional identity matrix.

After calculating the loadings and scores, the KPLS model is described as:

$$\begin{cases} \widehat{Y} = KU(T^T KU)^{-1}T^T Y \\ \widehat{Y_t} = K_t U(T^T KU)^{-1}T^T Y \end{cases} \tag{8}$$

where $\widehat{Y}$ presents the prediction outputs of the training samples, $\widehat{Y_t}$ is the prediction outputs of the testing samples, and K_t denotes the kernel matrix of the test samples.

3 Suggested RKPLS method

In many domains, large datasets have been presented. The training data, using the kernel method, has had a great success with an elegant treatment of data for monitoring systems. As a result, this number of observations must be stored in a memory.

Specifically, the KPLS technique presents a great disadvantage when the number of observations increases. This necessitates the use of large computer memory and training time [35]. Then, the learning time and CT for detection go up rapidly with the number of observations. Despite the fact that the KPLS technique solves the problem of nonlinearity, the calculation and memory problem that arise are posed for dynamic processes being monitored.

The dimensional kernel matrix (Gram matrix) is limited in this method [36, 37].

The main goal of the proposed RKPLS method is to reduce CT. From N measurement variables defined by the data matrix, we just choose the important observations. In this case, we obtain a parameter number of the kernel matrix equal to the L number of the selected latent components.

As a first step, we consider Eq. 9, which consists in the approach of the latent variables $\{w_j\}_{j=1..P}$ in the transformed input data $\phi\left(x_{Latent}^{(j)}\right) \in \phi\{x^i\}_{i=1...M}$ in order to get, as a second step, the highest projection value [29].

$$\phi\left(x_{Latent}^{(j)}\right) = \alpha_j * k_j(x), \qquad j = 1, 2..L \tag{9}$$

Generally, we can select the projection of all transformed data vectors $\phi\{x^i\}_{i=1...M}$ from the latent variables w_j to obtain the most loaded samples in terms of information $x_{Latent}^{(j)} \in \left\{x^{(i)}\right\}_{i=1...M}$, as depicted in Eqs. 10 and 11:

$$\begin{cases} \phi\left(x_{Latent}^{(j)}\right)_j = \max\limits_{i=1,..,M} \phi(x^i)_j \\ and \\ \phi\left(x_{Latent}^{(j)}\right)_{i \neq j} < \varsigma \end{cases} \tag{10}$$

where ς is a given threshold.

Furthermore, the RKPLS method consists in determining the reduced dataset by choosing the variables that have the highest projection variance in the direction of the selected latent components.

Then, the reduced data matrix of $\left\{x_{Latent}^{(j)}\right\}_{j=1..L}$ can be defined by Eq. 11:

$$X_r = \left[x_{Latent}^{(1)} \quad x_{Latent}^{(2)} \quad \dots \quad x_{Latent}^{(L)}\right]^T \tag{11}$$

Consequently, the reduced data matrix X_r given a reduced kernel matrix K_r related to the kernel function k and the number of the selected variables is indicated in Eq. 12:

$$K_r = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_L) \\ \vdots & \ddots & \vdots \\ k(x_L, x_1) & \dots & k(x_L, x_L) \end{bmatrix} \in R^{L \times L} \tag{12}$$

The main algorithmic steps of the suggested RKPLS are shown in Algorithm 2 as follows:

Algorithm 2 RKPLS algorithm

Input: N×M input data matrix X and N×L output data matrix Y

Output: Reduced input score matrices T, reduced output score matrix U

Step 1: Acquire an initial standardized block of training data $\{x_i\}_{i=1..N}$ and scale them,

Step 2: Construct the kernel matrix K and scale it,

Step 3: Project $\{\phi_i\}_{i=1..N}$ on the component latent $\{w_i\}$ and choose $x_{Latent}^{(i)}$ which satisfies Eq. 10,

Step 4: Construct the reduced kernel matrix $K_r \in R^{L \times L}$ following Eq. 12,

Step 5: Estimate the reduced KPLS model,

Step 6: Determine the control limits of the SPE chart present in the next section.

3.1 FD indices

The traditional PLS–based monitoring method uses, in general, the Hotelling's T^2 and the SPE or the Q-statistic, which are expressed respectively in terms of Mahalanobis and Euclidian distances [38, 39]. Therefore, the process monitoring step of the nonlinear PLS version (KPLS and the proposed RKPLS) is similar to that used in the PLS method. The T^2 statistic index is, usually, a way of measuring the projections of the observations in the feature space at various time samples [40]. T^2 is calculated as illustrated in Eq. 13:

$$T^2 = X^T \widehat{W} \widehat{\Lambda}^{-1} \widehat{W}^T X \tag{13}$$

where the diagonal matrix containing the eigenvalues is defined by

$\widehat{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_l)$, and $\widehat{W}$ is the weights matrix. The control limit is calculated for the T^2 index, utilizing the F-distribution, as follows:

$$T^2_\alpha = \frac{l(N-1)}{N-l} F_{l,N-l,\alpha} \tag{14}$$

where α is the significance level, l is the number of retained principal components, N is the number of observation in the dataset, and $F_{l,N-l}$ is the Fisher F distribution.

On the one hand, the SPE index allows FD in the residual subspace. The essential goal is to detect a new event for a new observation [41]. Furthermore, the SPE is computed as the sum of squares of the residuals by the KPLS model. To ensure the FD of the kernel method, the SPE is usually used in this step. Then, this index is characterized by the sensibility to model errors and also the addiction to the retained number.

The SPE, obtained from PLS, is given by Eq. 15:

$$SPE = \|X - \widehat{X}\|^2 = \|(I - \widehat{W}\widehat{W}^T)X\| \tag{15}$$

In this case, the confidence limit is presented in the Eq. 16. This index is determined using the χ^2 distribution. The process is considered abnormally functioning for the SPE index if:

$$SPE(k) > g\chi^2_{h,\alpha} \tag{16}$$

where g and h are presented respectively by $\frac{b}{2a}$ and $\frac{2a^2}{b}$. In this case, a is the estimated mean of the SPE and b is the variance of the SPE.

On the other hand, the EWMA can be applied to residues. The EWMA is used to improve the quality of the FD procedure and essentially reduce the FAR. This type of filter is considered to monitor variables in each data point. The EWMA chart presents many advantages. The performance of the EWMA is characterized by the ability to better detect small faults. Furthermore, the EWMA control limit consists in improving the detection abilities of small faults to the SPE chart [42]. The general expression of the EWMA applied to residues is given by:

$$Z_i = \lambda \bar{X}_i + (1-\lambda) Z_{i-1}, \quad i = 1..N \tag{17}$$

where λ is chosen such that $0 < \lambda \leqslant 1$, i is defined as the sample number, λ is defined as the smoothing parameter, $\bar{X}_i$ is defined by the average of the ith sample, and Z_{i-1} depends on the past information. The initial value is initialized following the average of the preliminary samples.

3.2 Flowchart of proposed method

To detail the principle of the RKPLS method, a flowchart is shown in Fig. 1. The flowchart represents the necessary steps of the FD for the suggested RKPLS.

4 Selection of kernel parameter using Tabu search algorithm

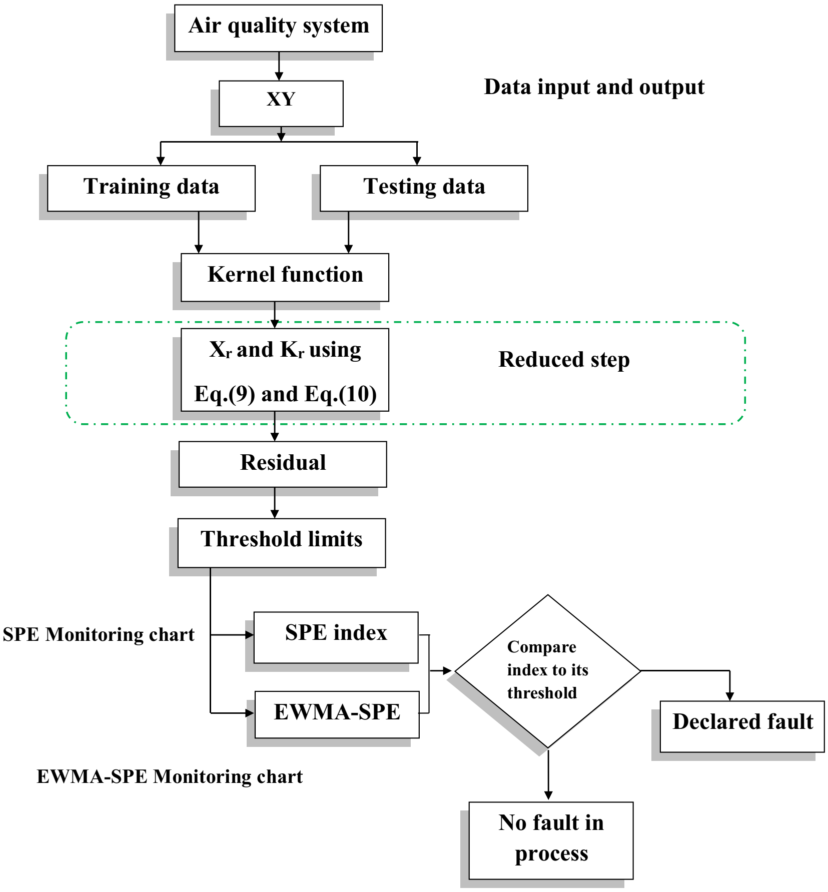
4.1 Principle

The kernel function is the core of the kernel method which helps it get an optimal solution. In general, the RBF kernel, as a nonlinear kernel function, is a reasonable first choice. Parameter σ is a key element of the RBF kernel and directly exerts considerable influence on the generalization ability of the KPLS method. The selection of the kernel function and the corresponding parameter are the key of KPLS. The σ parameter of the kernel function has an effect on the partitioning outcome in the feature space. If the value of σ is too large, it will lead to overfitting. If the value of σ is too small, it will lead to underfitting. In this part, we present an approach to select what is an optimal Gaussian kernel parameter to use when applying the proposed RKPLS technique. The optimal kernel parameter is defined as the one that can improve the fault detection performance. For many industrial applications, minimizing the false alarm rate may be the greatest performance criterion. Therefore, the choice of the Gaussian kernel parameter needs to be selected based on the given application. The Tabu search algorithm is applied to optimize the kernel parameter to use when applying the RKPLS algorithm.

4.2 Initial solution

In this study, the determination of the initial solution in the Tabu search algorithm is to optimize σ for the current RKPLS model. Firstly, an initialization solution is presented randomly. To reduce the search space referring

Fig. 1 Representation of RKPLS fault detection model



to the previous literature using the RKPLS model, it is recommended to introduce the constraints of parameter σ, which respectively attribute to the range $\sigma \in \left[2^{-6}, 2^{6}\right]$. The solution is computed by appending the nearest unused neighbor values of the parameter while improving the FD performance. The process is repeated until all the neighbors are visited.

5 Application

In this section, the RKPLS monitoring scheme is evaluated on the air quality network process.

5.1 Description of air quality monitoring

AIRLOR, the air quality monitoring network located in Lorraine (France), consists of 20 stations located in rural, peri-urban, and urban sites. Each monitoring station, for this model, consists of a set of sensors for measuring the concentrations of pollutants: carbon monoxide (CO), oxides of nitrogen (NO and NO_2) measured by the same analyzer, dioxide sulfur (SO_2), and ozone (O_3) [38, 43]. On the other hand, some stations (more precisely seven stations) are dedicated to the recording of additional meteorological parameters. Figure 2 shows an example of an air quality station.

Fig. 2 Air quality monitoring station

5.2 Air quality settings

The air quality, in general, is a secondary pollutant produced by complex photochemical reactions between primary pollutants, more precisely the nitrogen oxides NO, NO_2, and VOC emitted into the atmosphere [44, 45]. Then, the sensors, principally of the ozone concentration (O_3) and nitrogen oxides (NO and NO_2), monitor and detect the functioning abnormalities. On the one hand, O_3 is a secondary pollutant whose spatial distribution of the maximum values is rather homogeneous at our local scale. Whereas, the nitrogen oxides are primary pollutants which are more localized because their concentrations directly depend on the sources of emissions. Owing to its adverse health effects, tropospheric ozone has become one of the most studied topics in the recent decade. However, we can wonder about the performances of technical KPLS, which is the interest of the following section.

The air quality is produced through a complex series of reactions involving nitrogen oxides (NO_2 and NO) and volatile organic compounds (VOC) which are formed in the lower atmosphere by chemical reactions and secondary pollutants [44, 46].

In turn, each monitoring station contains a set of sensors dedicated to measuring the following concentrations of pollutants: CO (carbon monoxide), NO and NO_2 (oxides of nitrogen), SO_2, and O_3, respectively [1].

In this paper, we consider just six neighbor measurement stations. In this situation, matrix X contains 18 variables, respectively, named υ_1 to υ_{18}, of O_3 and nitrogen oxides (NO_2 and NO) collected from each station.

Furthermore, the essential purpose is to detect the functioning abnormalities of the sensors, principally those of the nitrogen oxides (NO and NO_2) and the ozone concentration O_3. Then, sensor faults, whose magnitude is approximately 20% of measurement for O_3, are simulated and 1080 samples are obtained.

To validate the two objective functions, the size and FAR, from the Tabu search algorithm, the optimal value of σ is equal to 25.37.

Nevertheless, the training data $X_{training}$ has 200 samples, among 1080, and the same choice for the testing data $X_{testing}$. For the next simulation results, two fault scenarios representing two different types of faults are generated to show the performance of the developed FD method.

- For the first example, fault 1 is a step bias of the sensor measuring ozone O_3 of variable υ_7. The fault is introduced between instances 50 and 100. Figures 3 and 4 show the detection index SPE for this fault.
- For the second example, the sensor measuring the nitrogen oxides NO_2, of the variable υ_{12}, is assumed to be faulty with a step bias representing the fault 2. The fault is introduced between instances 150 and 200.

5.3 Case 1: Fault in ozone O_3

As a first try, we present a fault in ozone O_3 for the station 3 in sample intervals of [50 to 100].

The FD results of the KPLS-based SPE and KPLS-based EWMA-SPE techniques are depicted in Fig. 3.

The FD results of the proposed RKPLS-based SPE and RKPLS-based EWMA-SPE techniques are illustrated in Fig. 4. More precisely, these figures represent, respectively, the time evolution of the SPE different indices and the evolution of the different EWMA-SPE indices.

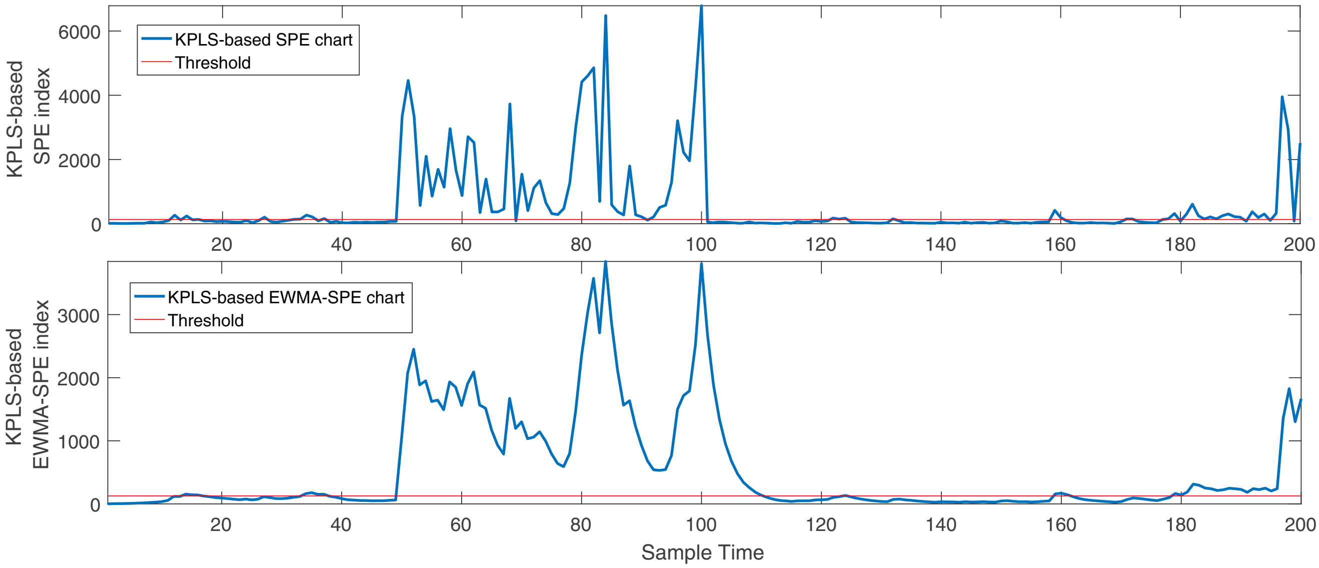


Fig. 3 Monitoring faults in ozone O_3 using KPLS in sample intervals of [50 to 100]

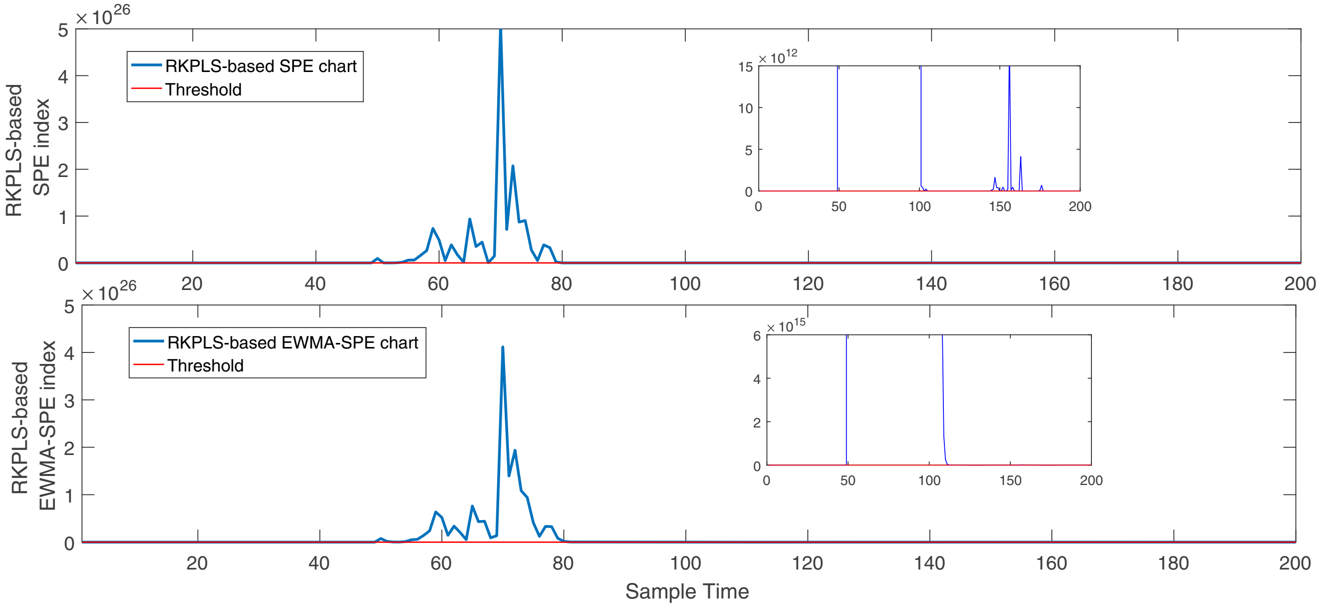


Fig. 4 Monitoring faults in ozone O_3 using proposed RKPLS in sample intervals of [50 to 100]

The compared performances of the suggested RKPLS for the air quality system in terms of FAR, GDR, and CT are summarized in Table 2.

The proposed RKPLS provides a reduced kernel matrix with 44 observations. According to Table 2, the suggested RKPLS method has a less FAR compared to the classical KPLS. To show the performance of the proposed RKPLS methods, we determine the CT of the different methods. However, the RKPLS-based SPE and RKPLS-based EWMA-SPE techniques show better FD results compared to the conventional KPLS-based SPE and KPLS-based EWMA-SPE techniques for the test of faults in ozone O_3, as illustrated in Figs. 3 and 4.

5.4 Case 2: Fault in nitrogen oxides (NO_2)

Here, we present a fault in the nitrogen oxides NO_2 for station 4 in sample intervals of [150 to 200]. To evaluate the obtained results, Fig. 5 depicts the FD results of the KPLS-based SPE and KPLS-based EWMA-SPE techniques.

Afterwards, Fig. 6 shows the FD results of the proposed RKPLS-based SPE and RKPLS-based EWMA-SPE techniques. These figures represent, respectively, the time evolution of the different SPE indices and the evolution of the different EWMA-SPE indices.

The detection results out of both simulated methods, KPLS and RKPLS, using the SPE and EWMA-SPE statistics in the failure condition are provided in Table 3.

The result of the application of the suggested RKPLS method to the air quality process is demonstrated in Figs. 5 and 6. For this test, the FD results show also that the suggested RKPLS technique gives a GDR for a single fault in the nitrogen oxides NO_2 compared to the KPLS with some FAR.

Furthermore, Table 3 shows that the proposed RKPLS-based SPE and RKPLS-based EWMA-SPE techniques provide better results compared to both conventional KPLS-based SPE and EWMA-SPE.

Finally, the reduced technique based on SPE and EWMA-SPE shows better fault detection performances

Table 2 Summary of good detection rates, false alarm rates, and computation time for TEP data for case 1: fault in ozone O_3

Chart/fault detection metric	FAR (%)	GDR (%)	CT (s)
KPLS-based SPE	17.63	84	0.2387
KPLS-based EWMA-SPE	14	100	0.2387
RKPLS-based SPE	8.66	100	0.1348
RKPLS-based EWMA-SPE	4	100	0.1348

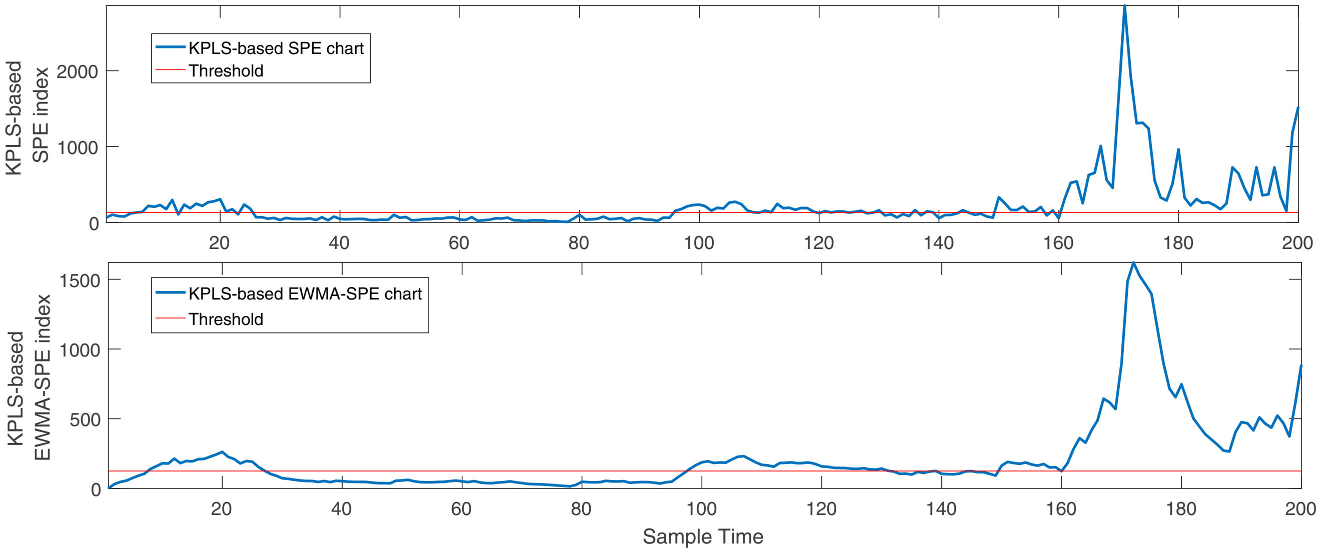


Fig. 5 Monitoring faults in the nitrogen oxides NO_2 using KPLS in sample intervals of [150 to 200]

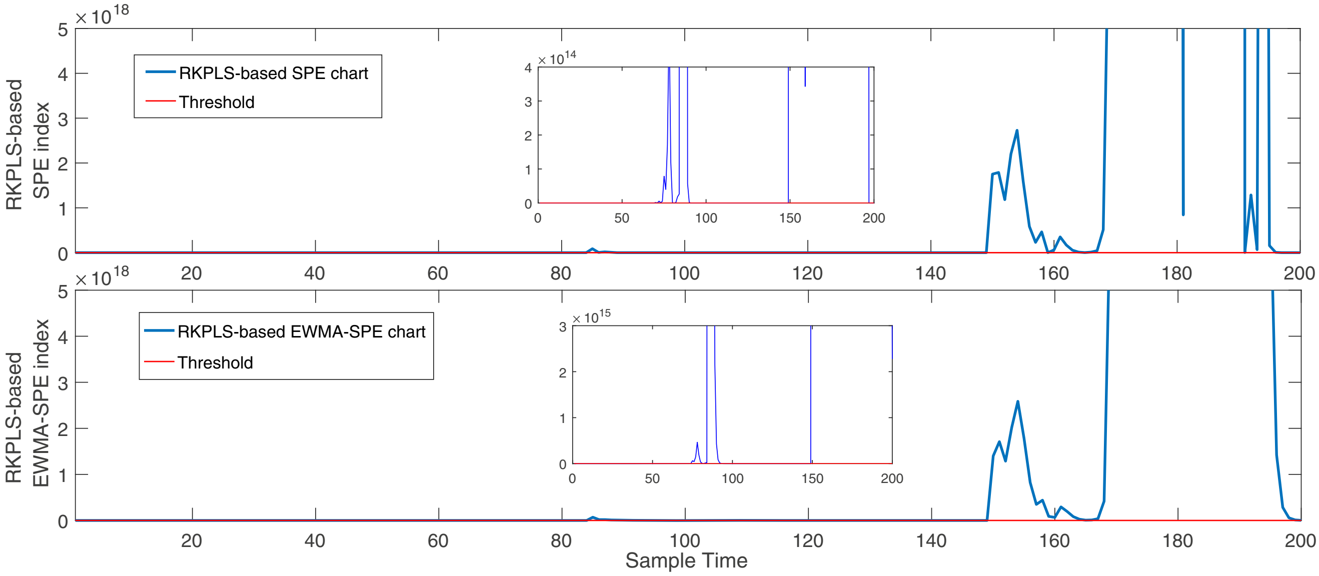


Fig. 6 Monitoring a faults in the nitrogen oxides NO_2 using proposed RKPLS in sample intervals of [150 to 200]

Table 3 Summary of good detection rates, false alarm rates, and computation time for TEP data for case 2: fault in nitrogen oxides NO_2

Chart/fault detection metric	FAR (%)	GDR (%)	CT (s)
KPLS-based SPE	17.67	94	0.2263
KPLS-based EWMA-SPE	15.33	95.58	0.2263
RKPLS-based SPE	10.66	97.9	0.1368
RKPLS-based EWMA-SPE	5.33	100	0.1368

with FAR, GDR, and also CT, as illustrated in Tables 2 and 3.

6 Conclusion

To avoid the high levels of air pollution, in this paper, we have proposed a new fault detection method applicable to the air quality process monitoring using the RKPLS technique. To control the pollution rate and protect human health, a reduced KPLS method has been opted for. The purpose of this suggested method is to detect the functioning abnormalities of the sensors, principally those of the ozone concentration (O_3) and nitrogen oxides (NO and NO_2). Afterwards, a reduced optimized RKPLS technique has been suggested in order to extend the advantages of the KPLS models to the air quality process.

This paper aims to improve RKPLS for fault detection. Nevertheless, the RKPLS-based SPE and RKPLS-based EWMA-SPE fault detection performances are assessed and compared to those of the classical KPLS-based SPE. Firstly, an optimal and reduced kernel parameter using the important latent components has been selected in order to enhance the use of the classical KPLS model. Secondly, the developed detection method has utilized, at the first time, the different indices of the SPE and then has combined the various indices of the SPE and EWMA to detect and modify the average residual of the air quality model. The developed RKPLS-based SPE method has shown improved FD at the level of FAR, GDR, and the CT, mostly, when compared to the KPLS-based SPE. However, using the EMWA, we get better FD results in all cases, compared to the methods based on the SPE. The RKPLS-based EMWA-SPE technique has indicated a slightly weak FAR. Furthermore, the performance of the RKPLS method is good compared to KPLS. To do that, the results have demonstrated the efficiency of the developed technique in terms of FAR, GDR, and CT compared with the conventional fault detection KPLS. The relevance of the suggested monitoring technique has been shown for fault detection on air quality monitoring network data.

In fact, the RKPLS method is suggested in a static version. For a dynamic nonlinear system, an online version can be proposed.

References

1. Harkat MF, Mourot G, Gilles R (2006) An improved PCA scheme for sensor FDI: application to an air quality monitoring network. J Process Control 16(6):625–634
2. Mofarrah A, Husain T (2010) A holistic approach for optimal design of air quality monitoring network expansion in an urban area. Atmos Environ 44(3):432–440
3. Yang Z, Wang J (2017) A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction. Environ Res 158:105–117
4. Liu MK, Avrin J, Pollack R, Behar J, McElroy J (1986) Methodology for designing air quality monitoring networks: I. theoretical aspects. Environ Monit Assess 6(1):1–11
5. Stanimirova I, Simeonov V (2005) Modeling of environmental four-way data from air quality control. Chemom Intell Lab Syst 77(1-2):115–121
6. Zheng J, Zhong L, Wang T, Louie P, Li Z (2010) Ground-level ozone in the pearl river delta region: analysis of data from a recently established regional air quality monitoring network. Atmos Environ 44(6):814–823
7. Jaffel I, Taouali O, Harkat MF, Messaoud H (2015) Online process monitoring using a new PCMD index. Int J Adv Manuf Technol 80(5-8):947–957
8. Willsky A, Chow E, Gershwin S, Greene C, Houpt P, Kurkjian A (1980) Dynamic model-based techniques for the detection of incidents on freeways. IEEE Trans Autom Control 25(3):347–360
9. Venkatasubramanian V, Rengaswamy R, Gershwin S, Kavuri S, Yin K (2003) A review of process fault detection and diagnosis: Part III: process history based methods. Comput Chem Eng 27(3):327–346
10. Yan S, Huang J, Yan X, Kavuri S, Yin K (2003) Monitoring of quality-relevant and quality-irrelevant blocks with characteristic-similar variables based on self-organizing map and kernel approaches. J Process Control 73:103–112
11. Benothman K, Maquin D, Ragot R, Benrejeb M (2007) Diagnosis of uncertain linear systems: an interval approach. Int J Sci Tech Automatic Control Comput Eng 1(2):136–154
12. Lahdhiri H, Taouali O, Elaissi I, Jaffel I, Harakat MF, Messaoud H (2017) A new fault detection index based on Mahalanobis distance and kernel method. Int J Adv Manuf Technol 91(5-8):2799–2809
13. Joe Qin S (2003) Statistical process monitoring: basics and beyond. J Chemom 17(8-9):480–502
14. Jaffel I, Taouali O, Elaissi I, Jaffel I, Messaoud H (2013) A new online fault detection method based on PCA technique. IMA J Math Control Inf 31(4):487–499
15. Said M, Fazai R, Abdellafou K, Taouali O (2018) Decentralized fault detection and isolation using bond graph and PCA methods. Int J Adv Manuf Technol 99(1-4):517–529
16. Kano M, Tanaka S, Hasebe S, Hashimoto I, Ohno H (2003) Monitoring independent components for fault detection. AIChE J 49(4):969–976
17. Li G, Qin S, Zhou D, Hashimoto I, Ohno H (2003) Geometric properties of partial least squares for process monitoring. Automatica 46(1):204–210
18. Wold H (1985) Partial least squares. Encyclopedia of statistical sciences
19. Neffati S, Abdellafou K, Taouali O, Bouzrara K (2019) A new bio-CAD system based on the optimized KPCA for relevant feature selection. Int J Adv Manuf Technol: 1–12. https://doi.org/10.1007/s00170-018-03266-w
20. Harkat MF, Mansouri M, Nounou M, Nounou H (2018) Enhanced data validation strategy of air quality monitoring network. Environ Res 160:183–194
21. Tang J, Zhang J, Wu Z, Liu Z, Chai T, Yu W (2017) Modeling collinear data using double-layer GA-based selective ensemble kernel partial least squares algorithm. Automatica 219:248–262
22. MacGregor JF, Jaeckle C, Kiparissides C, Koutoudi M (1994) Process monitoring and diagnosis by multiblock PLS methods. AIChE J 40(5):826–838
23. Helland K, Berntsen HE, Borgen OS, Martens H (1992) Recursive algorithm for partial least squares regression. Chemom Intell Lab Syst 14(1-3):129–137

24. Zhou D, Li G, Qin SJ (2010) Total projection to latent structures for process monitoring. AIChE J 56(1):168–178
25. Rosipal R, Trejo LJ (2001) Kernel partial least squares regression in reproducing kernel hilbert space. J Mach Learn Res 2(Dec):97–123
26. Zhang Y, Du W, Fan Y, Zhang L (2015) Process fault detection using directional kernel partial least squares. Ind Eng Chem Res 54(9):2509–2518
27. Zhang Y, Hu Z (2011) Multivariate process monitoring and analysis based on multi-scale KPLS. Chem Eng Res Des 89(12):2667–2678
28. Kim K, Lee JM, Lee IB (2005) A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. Chemom Intell Lab Syst 79(1-2):22–30
29. Taouali O, Elaissi I, Messaoud H (2015) Dimensionality reduction of RKHS model parameters. ISA Trans 57:205–210
30. Willis A (2010) Condition monitoring of centrifuge vibrations using kernel PLS. Comput Chem Eng 34(3):349–353
31. Wang G, Jiao J, Yin S (2018) Efficient nonlinear fault diagnosis based on kernel sample equivalent replacement. IEEE Trans Ind Inf 3
32. Wang Q (2012) Kernel principal component analysis and its applications in face recognition and active shape models. arXiv:1207.3538
33. Lindgren F, Geladi P, Wold S (1993) The kernel algorithm for PLS. J Chemom 7(1):45–59
34. Rosipal R, Geladi P, Wold S (2010) Nonlinear partial least squares: an overview. Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques: 169–189
35. Jaffel I, Taouali O, Harkat MF, Messaoud H (2017) Kernel principal component analysis with reduced complexity for nonlinear dynamic process monitoring. Int J Adv Manuf Technol 88(9-12):3265–3279
36. Taouali O, Jaffel I, Lahdhiri H, Harkat MF, Messaoud H (2016) New fault detection method based on reduced kernel principal component analysis (RKPCA). Int J Adv Manuf Technol 85(5-8):1547–1552
37. Lahdhiri H, Taouali O, Elaissi I, Harkat MF, Messaoud H (2018) Nonlinear process monitoring based on new reduced Rank-KPCA method. Stoch Env Res Risk A 32(6):1833–1848
38. Lahdhiri H, Said M, Abdellafou K, Taouali O, Harkat MF, Messaoud H (2019) Supervised process monitoring and fault diagnosis based on machine learning methods. Int J Adv Manuf Technol (1–17)
39. Liu X, Kruger U, Elaissi I, Littler T, Xie L, Wang S (2009) Moving window kernel PCA for adaptive monitoring of nonlinear processes. Chemom Intell Lab Syst 96(2):132–143
40. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417
41. Jackson JE, Mudholkar GS (1979) Control procedures for residuals associated with principal component analysis. JTechnometrics 21(3):341–349
42. Lee C, Choi SW, Lee I (2004) Sensor fault identification based on time-lagged PCA in dynamic processes. Chemom Intell Lab Syst 70(2):165–178
43. Fazai R, Abdellafou K, Said M, Taouali O (2018) Online fault detection and isolation of an AIR quality monitoring network based on machine learning and metaheuristic methods. Int J Adv Manuf Technol: 1–14
44. Bell ML, McDermott A, Zeger SL, Samet JM, Dominici F (2004) Ozone and short-term mortality in 95 US urban communities, 1987-2000. Jama 292(19):2372–2378
45. Harakat MF, Mourot G, Ragot J (2009) Multiple sensor fault detection and isolation of an air quality monitoring network using RBF-NLPCA model. IFAC Proceedings 42(8):828–833
46. Zhang T (2001) An introduction to support vector machines and other kernel-based learning methods. AI Mag 22(2):103
47. Qin SJ (2012) Survey on data-driven industrial process monitoring and diagnosis. Annu Rev Control 36(2):220–234
48. Jalali-Heravi M, Kyani A (2007) Application of genetic algorithm-kernel partial least square as a novel nonlinear feature selection method: activity of carbonic anhydrase II inhibitors. Eur J Med Chem 42(5):649–659