The Report committee for Ray Chen Wang

Certifies that this is the approved version of the following report:

Visualization of Multivariate Process Data for Fault Detection and Diagnosis

APPROVED BY SUPERVISING COMMITTEE:

Supervisor: _____

Michael Baldea

Co-Supervisor:_____

Thomas F. Edgar

Visualization of Multivariate Process Data for Fault Detection and Diagnosis

by

Ray Chen Wang, B.S.

Report

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2014

Visualization of Multivariate Process Data for Fault Detection and Diagnosis

by

Ray Chen Wang, MSE

The University of Texas at Austin, 2014

SUPERVISOR: Michael Baldea, Thomas F. Edgar

This report introduces the concept of three-dimensional (3D) radial plots for the visualization of multivariate large scale datasets in plant operations. A key concept of this representation of data is the introduction of time as the third dimension in a two dimensional radial plot, which allows for the display of time series data in any number of process variables. This report shows the ability of 3D radial plots to conduct systemic fault detection and classification in chemical processes through the use of confidence ellipses, which capture the desired operating region of process variables during a defined period of steady-state operation. Principal component analysis (PCA) is incorporated into the method to reduce multivariate interactions and the dimensionality of the data. The method is applied to two case studies with systemic faults present (compressor surge and column flooding) as well as data obtained from the Tennessee Eastman simulator, which contained localized faults. Fault classification using the interior angles of the radial plots is also demonstrated in the paper.

Table of Contents

Chapter 1: Introduction

Thanks to the ubiquitous use of technology in today's society, there has been a substantial increase in the availability of large-scale datasets and "Big Data". These datasets are present in many different systems such as social media, the finance sector, and in manufacturing or process industries. The advent of data transmission technology and cheaper storage capabilities have pushed the generation of more granular and varied data, particularly in the chemical process industries.

A downside to these large datasets is that they are hard to analyze and cumbersome to manipulate, so extracting information from the data, or data mining, is difficult. Visualization of the data is often the first step to understanding and making sense of the data, so how the data is represented visually is crucial.

Most process plants today are outfitted with a large variety of sensors that monitor and maintain the proper operation of a plant. Measurements from these sensors are taken often at high sample rates as well, so both the dimensionality as well as the sample size of these measurement data is large. One of the many uses of this sensor data is to calculate performance parameters and use these parameters to identify process faults. Many fault detection methods such as principal component analysis (PCA) and partial least squares (PLS) regression exist and have been used to detect and classify faults in both industrial case studies and *in silico* simulators [1–3].

Further benefits can be obtained if systematic faults, as opposed to individual component faults or breakdowns can be detected, if not predicted ahead of a time. Two

examples of systemic faults looked at in this report include compressor surge and column flooding.

In this report, a new representation of large, multivariate datasets for visualization purposes is proposed. Through this method large datasets can be visualized appropriately and any trends present in the data can be identified. This data representation method is further explored for use in fault detection and classification. By taking into account data collected by all sensors and using the proposed multivariate data representation method, systematic fault detection as well as prediction can be done on industrial datasets.

This paper will introduce the concept of 3D radial plots as well as mechanisms for fault detection and classification based on the 3D radial plot. We also present case studies of industrial datasets that have been analyzed using the presented method. Simulator data from the Tennessee Eastman Process Simulator is also used as a benchmarking tool for comparison with other methods.

Chapter 2: Preliminaries

2.1 Representations of Multi-Dimensional Data

There are many different ways to represent data, the most common of which are score plots. Score plots display 2 variables and plot them against one another. Using these plots it is easy to understand how a variable evolves in time by using time as one of the variables plotted. Its downside is that multiple score plots are needed to monitor different variables, which is difficult for plant operators to keep track of. While score plots are explicit in time, the number of variables that can be displayed is severely limited.

To lessen this particular problem, the addition of one more coordinate can be added to construct a 3D Cartesian plot, where 3 variables can be displayed at a time. This reduces the number of plots needed to display all the data, but the time-series nature of the process data is lost – it is impossible to determine which sample of data came first. The 3D Cartesian plot is a time implicit representation of data.

Parallel coordinates is a method for presenting multivariate data on a single plot and was first proposed by Inselberg [4]. This is done by giving each variable a vertical axis and arranging the axes in parallel. Each sample is an open line that connects the sample values across the different axes. The number of axes can be increased indefinitely and is only limited by screen size and display resolution [5,6].

However, parallel coordinates also suffer from cluttering issues. This is especially true for large datasets with many samples – the sheer number of lines clutters the plot and makes it difficult to discern any meaning information. Interacting with the plot through brushing or tiling

methods are needed to mitigate this issue [7–9]. Parallel coordinates is also a time implicit data representation method since time is considered just one of many parallel coordinates, which belies its importance for fault detection. Therefore it is desirable to have a multivariate, yet time explicit representation of data.

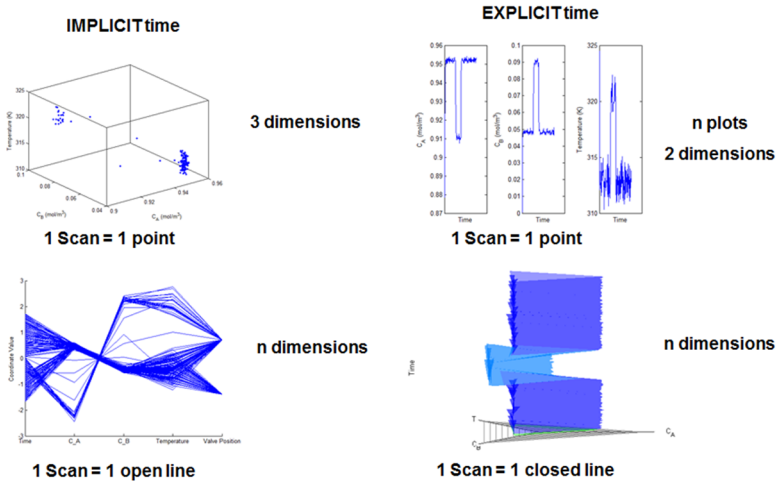


Fig 1. Commonly known methods of plotting

From top left, in clockwise order: Cartesian 3D plots, score plots, 3D radial plots (proposed), parallel coordinate plots

Chapter 3: Data Representation Using Multi-Dimensional Radial Plots

This section introduces the proposed data representation as well as the related fault detection and classification mechanisms. A sample set of data, which is referred to as DS1, is used to introduce the theoretical concepts and is available in the appendix.

3.1 Radial Plots

Radial plots, also known as Kiviat diagrams or spider plots, are a variation of the parallel coordinates plot, where parallel axes are wrapped radially around a center point. The time dimension is not included as one of the radial axes and is instead included as an axis normal to the plotting plane at the center point [10].

3.2 Geometric Construction of Multi-Dimensional Radial Plots

In constructing multi-dimensional radial plots, we rely on data that is normalized and mean-centered – each variable has zero mean and unit standard deviation.

$$ScaledData_k = \frac{OriginalData_k - mean_k}{StdDev_k} \quad \text{(Eq. 1)}$$

Where the subscript *k* indicates the k^{th} variable in the dataset.

The number of axes *N* to plot is a user-defined parameter, taking into consideration issues with plot resolution – too many axes make it difficult to visualize the data – and the number of variables in the dataset.

First, the position of the axes is calculated in polar coordinates, beginning with the first axis at 0 radians. Subsequent axes are spaced $2\pi/N$ radians apart. To determine the overall

radius of the plots, the maximum and minimum values of each axis across the entire dataset are found. The maximum and minimum of those values, which are called the overall maximum and minimum, are then used to determine the radius of the entire radial plot.

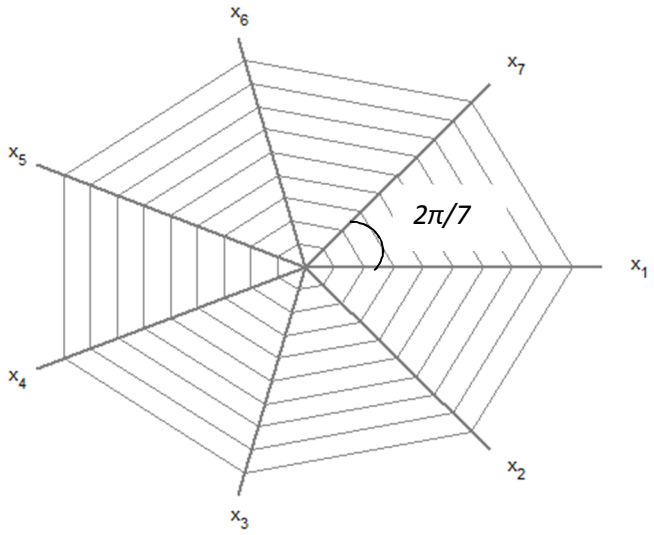
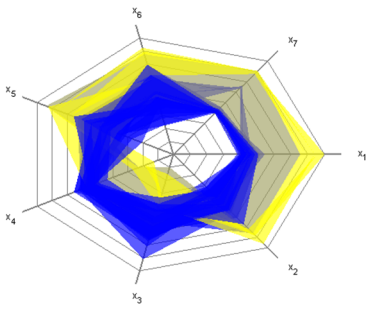


Fig 2. Radial plot shape

To plot the scaled data points, they need to be further adjusted to represent trends in the data accurately. The overall minimum of the respective variable is subtracted from each data point; the resulting value is then divided by the difference between the overall maximum and minimum, obtaining the radial position of each point. A fractional gain is used to ensure that the data points stay within the bounds of the radial plot, and a bias is added to ensure that the minimum points on each axis do not cross into the opposing axes on the other side of the origin [11]. Equation 2 provides the mathematical form of the scaling method.

$$Radial\ position = \frac{gain*(data\ point - overall\ minimum)}{(overall\ maximum - overall\ minimum)} + bias \qquad \text{(Eq. 2)}$$

The time dimension is captured via an axis normal to the plane of the radial plot axes passing through the origin. Each closed line or polygon connects the sample values between each axis to form one sample of the dataset. These "data slices" can then be stacked on top of one another in time to form a 3D figure that resembles a cylinder, with time as the *z*- axis, with each slice corresponding to a sample of the dataset in time.



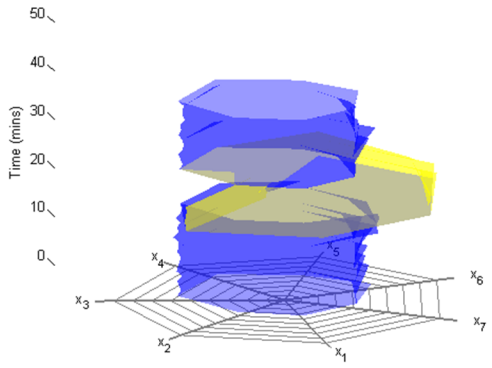


Fig 3a. Radial plots in 2D using DS1 **Fig 3b.** Radial plots in 3D using DS1

These 3D radial plots can be updated in real time – as measurements from the process are streamed from different sensors, additional "data slices" can be stacked on top of the current figure to show the current state of the plant. New data can be added at the top of the plot while old data can be removed from the visualization in a "first-in-first-out" fashion. The amount of data to display at any point is user-defined and ideally dependent on the time scale of the dataset and resolution desired.

3.3 Representing Large Datasets and Avoiding Cluttering Effects

Graphical representations of data can be affected by cluttering problems, particular for large datasets. Cluttering refers to the visual impact of the overlapping of the plots due to the

large number of samples, which reduces the clarity of the plot and leads to the loss of information and insight. Parallel coordinate plots suffer from such an issue due to large amounts of data because the number of lines that need to be plotted becomes too many, obscuring any individual line from being identified. Radial plots suffer from a similar issue but for a different reason – instead of too much data, too many variables results in cluttering problems. This is a resolution concern: Only 360 degrees are available for use as plotting, so plotting a large number of variables makes it difficult to distinguish between individual variables without adequate spacing.

This issue can be addressed in several ways. Assuming that sufficient expert process knowledge is available, the subset of data to be plotted can be user-selected. However, this approach is less practical when dealing with an unfamiliar process, so a more rigorous process for reducing the dimensionality of the dataset should be used.

PCA will be used to reduce the dimensionality of the data – this is done since PCA transforms combination of variables into principal components, each of which capture a certain amount of variance in the dataset. These principal components can be selected based on the total variance captured by each principal component, with those capturing more variance being favored for plotting. PCA has been used extensively in literature to conduct fault detection and classification in various processes, and excellent reviews on its use are available [12–16].

After applying PCA to the data, the PCA scores (instead of the original data), will be used to construct the 3D radial plots. Depending on the level of variance desired to be captured by the principal components, the number of axes on the resulting plot is likely lower than the

corresponding plot of the original data. Alternatively, the number of axes on the plot can be fixed and the variance captured by the principal components allowed to vary.

For easy standardization of the visualizations, PCA was used for all analyses in the paper and the number of principal components to plot was fixed at five.

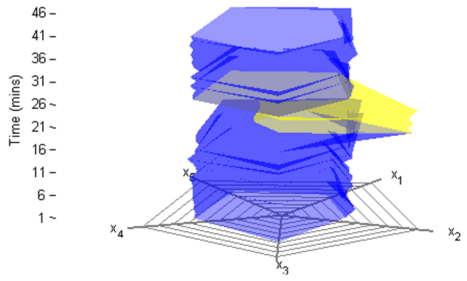


Fig 4. Radial plots in 3D with five principal components selected (data in DS1 represented)

3.4 3D Radial Plots as Multivariate Control Charts

One of the more important contributions of analyzing process data is obtaining information concerning potential faulty operating conditions. Fault *detection* focuses on determining that a process fault has occurred; the exact type of fault can be determined through a fault *isolation* mechanism. The following subsections will introduce the fault detection and classification mechanism to be used in the proposed radial plots framework.

3.4.1 Fault Detection

Fault detection is an important area of process monitoring as the cost of recovery and repair of a plant from a faulty state increase with process complexity. Many methods of fault detection exist, and can be categorized into different groups: *Data-based methods* such as PCA, and spectral analysis extract information from process operating data. *Model-based methods*

usually rely on building a first principles model of the plant and then comparing data from the plant to the model to determine if a fault has occurred. *Knowledge-based methods* use expert knowledge from plant operators to define rules that help determine if a faulty condition has occurred or not. There are several papers available that extensively review different methods in each category, and such a review is beyond the scope of this report [3,17–19].

Traditional fault detection methods account for parametric or operator faults, and may fail in the presence of a *systemic* malfunction, where a process-wide failure condition is reached, but no individual sensor indicate a faulty state. The two case studies looked at in the report – compressor surge and column flooding – are two such failure conditions.

3.4.2 Defining the "Normal" Operating State of a Process

Defining a fault-free, normal operating region is key to any fault detection mechanism. In this report, it is desirable to detect systemic malfunctions – the assumption is made that such faults manifest themselves as a deviation of the process from its nominal steady state. A moving window variance method is used in conjunction with a threshold to determine the steady state operation of the process [20]. The length of the moving window to use can be defined as some function of process time constants or sample rates. The threshold is set using the mean of the variance across the dataset. All data points with a variance below this threshold are considered to be potential steady state regions. The longest of these regions is then considered as the steady state region.

3.4.3 Constructing Multivariate Control Charts

The radial plots approach can function as a multivariate control chart. MacGregor and Kourti [15] have shown that the use of univariate control charts over multiple variables may result in "blind spots" where a process falls outside the joint normal operating region, but still resides within the individual control limits for each variable – an abnormal event would not be detected using univariate control charts in this case.

To use radial plots as a multivariate control chart, the concept of a centroid is first introduced. Geometrically, a centroid is the average of all bisectors of vertices in a polygon. For a regular polygon, the centroid would be in the center of the polygon. Since data is always normalized prior to plotting, the steady state region of the data would consist of near-regular polygons centered around the origin of the radial plot. This means that the centroid of the steady state region is clustered near the origin, with deviations occurring due to noise.

In the case of a fault, process variables change significantly and the regular polygon is warped into a non-regular polygon, resulting in a movement of the centroid away from the centroid. This movement is shown in Figure 5a.

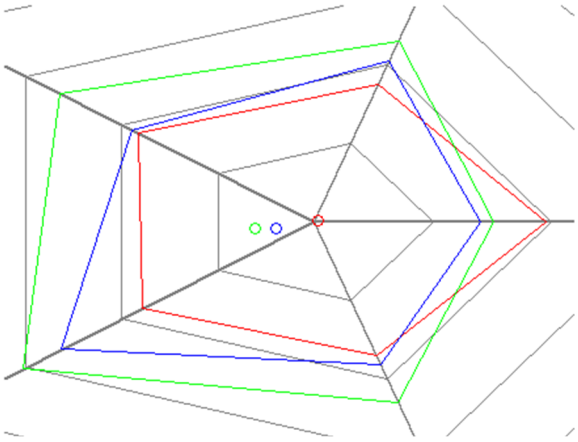


Fig 5a: Shift in centroids due to deformation of regular polygon (red)

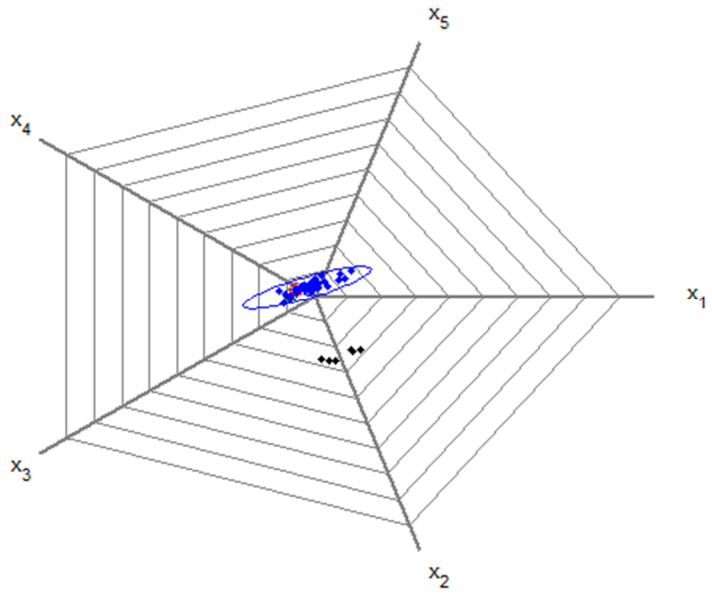


Fig 5b: Centroid plot of DS1 data with confidence ellipse

In Figure 5b, the confidence region for DS1 can be represented as an ellipse, with the faulty region having centroids (in black) that fall outside the confidence region.

The confidence ellipse is defined around the steady state region based on the centroids included in the steady state region. To do so, eigenvectors are used to transform a unit circle

into the desired ellipse. These eigenvectors are obtained from the covariance matrix of X and Y coordinates of the steady state centroids. A confidence interval can be specified by scaling the covariance matrix appropriately. For all cases in the paper the confidence interval used is 95%. The algorithm for doing so is described below:

$$\boldsymbol{c} = \boldsymbol{F^{-1}}(0.99 \mid 2), F \text{ is the CDF of the } \chi^2 \text{ distribution} \quad \text{(Eq. 3)}$$

$$[\boldsymbol{\lambda_1 v_1} \quad \boldsymbol{\lambda_2 v_2}] = \boldsymbol{eig}(\boldsymbol{c} * \boldsymbol{Cov}(\boldsymbol{Centroid\ X}, \boldsymbol{Centroid\ Y})) \quad \text{(Eq. 4)}$$

$$\begin{bmatrix} \boldsymbol{Ellipse\ X} \\ \boldsymbol{Ellipse\ Y} \end{bmatrix} = [\boldsymbol{\lambda_1 v_1} \quad \boldsymbol{\lambda_2 v_2}] * \begin{bmatrix} \boldsymbol{Unit\ Circle\ X} \\ \boldsymbol{Unit\ Circle\ Y} \end{bmatrix} \quad \text{(Eq. 5)}$$

Fault detection is done by reversing the process and obtaining X and Y coordinates of the tested point with respect to the unit circle (as opposed to the ellipse). A simple test can then be done to determine if the tested point lies within or outside the unit circle. If it lies outside the unit circle it can be considered a fault.

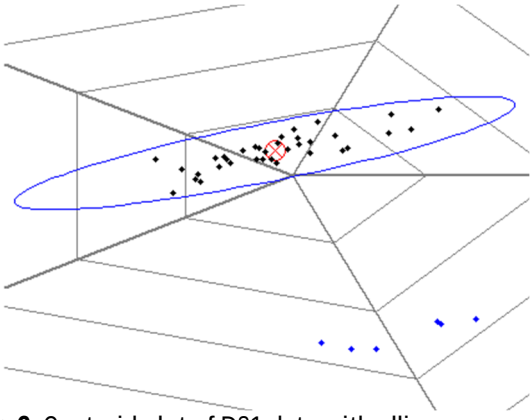


Fig 6. Centroid plot of DS1 data with ellipse
Points outside the ellipse (in blue) are considered faults

As seen in Figure 6, the steady state region (in black) has an ellipse defined around it and the points outside the ellipse (in blue) are determined as faults in the DS1 dataset.

Chapter 4: Fault Classification Based on Fault Signatures

The regular polygonal shape of a steady state data sample is distorted under the occurrence of a fault. It is expected that the shape of the distortion be the same for faults that are similar i.e. caused by the same phenomena or sequence of events. In order to quantify this similarity, we make use of the interior angles of the polygons of each data slice. Each polygon is deconstructed into many triangles as defined by the axes of the radial plot. The law of cosines is used to determine the interior angles. The values in the equations below refer to the values in Figure 7.

$$\cos(\alpha) = \frac{d_{51}^2 + d_{5o}^2 - d_{1o}^2}{2 d_{51} d_{5o}} \qquad \text{(Eq. 6)}$$

$$\cos(\beta) = \frac{d_{51}^2 + d_{1o}^2 - d_{5o}^2}{2 d_{51} d_{1o}} \qquad \text{(Eq. 7)}$$

Where d_{ij} is the distance between points i and j. The subscript *o* indicates the origin point (0,0).

After angles for each triangle are calculated, adjacent angles can be summed up to obtain the interior angles of the polygon.

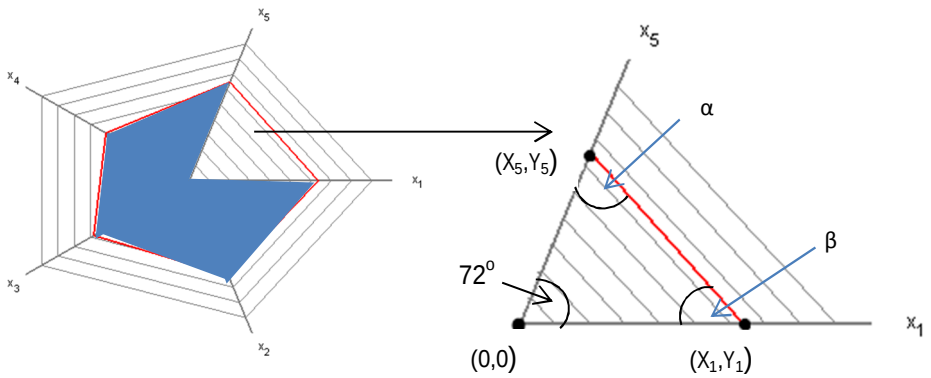


Fig 7. Computation of interior angles for a data slice
α *and* β *can be found using Eq. 6 and Eq.7*

Fault signatures are defined by the vector of the interior angles of the radial plots and the time evolution of said faults. Fault classification can be done on faults that exhibit the same or similar signature. Well-established data analysis methods such as dynamic time warping can be used in order to superpose/compare the plots of interior angles with time-displaced faults [21]. This is useful to compare if streaming data contains a fault belonging to a known fault class.

Consider the example of DS1, where a fault is present at t = 21 mins.

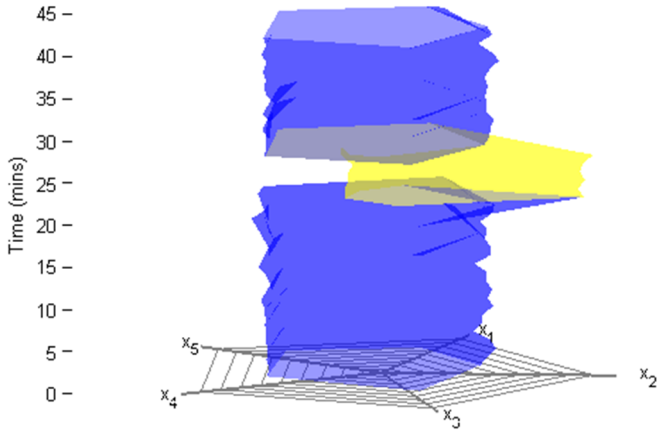


Fig 8. Dataset DS1 in radial plot representation
Yellow region indicates the problematic, potentially faulty region

A new dataset DS1' is generated to illustrate the fault classification procedure described above. DS1' has the faulty region in DS1 translated to a later time. PCA was also applied to the data and the resulting angle plots are presented below.

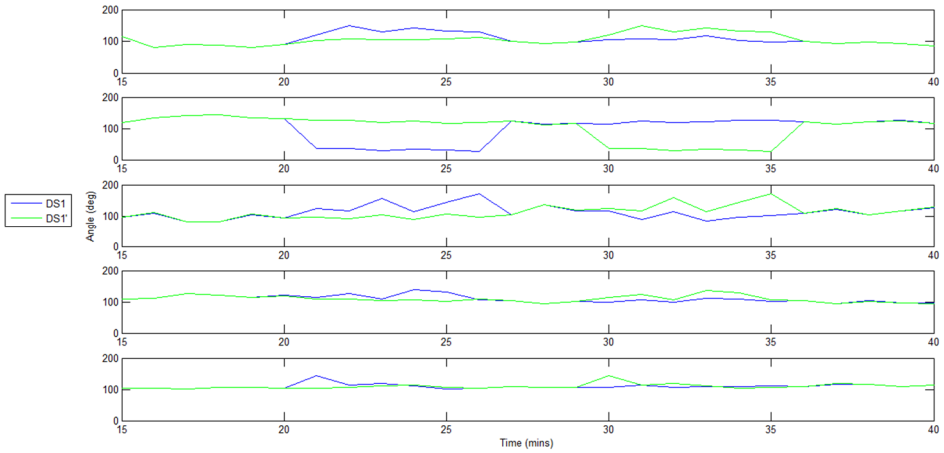


Fig 9a. Plot of all 5 interior angles of DS1 and DS1' dataset

The evolution of the interior angles for DS1 and DS1' is shown in Figure 9a. It is clear that potentially similar faults are occurring at different times. Dynamic time warping was used to compare the two datasets and the results are shown in Figure 9b.

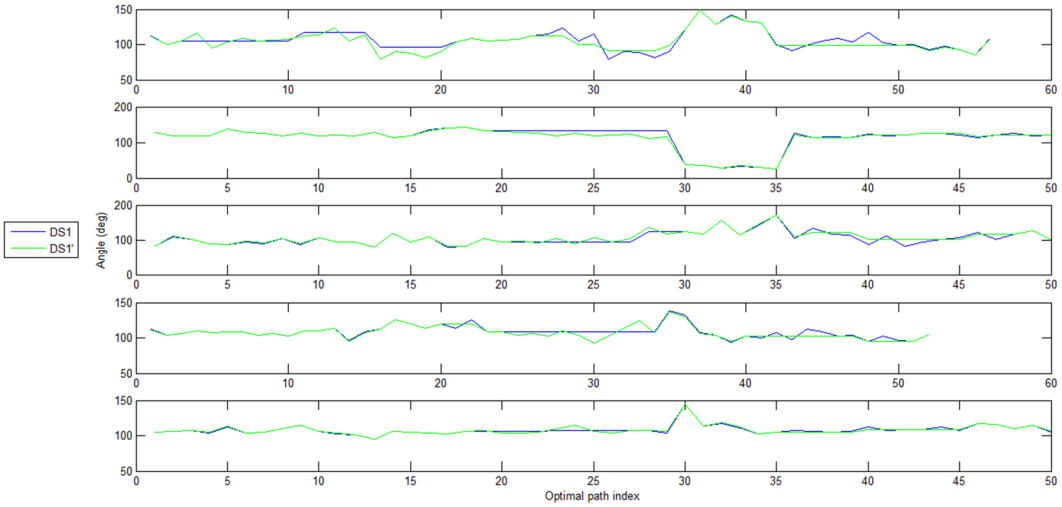


Fig 9b. Plot of all 5 interior angles of DS1 and DS1' dataset with dynamic time warping

It is clear that the two faults are very similar, if not the same. Further dynamic time warping results are shown later when exploring the compressor surge case study.

Chapter 5: Case Studies

5.1 Compressor Surge

Compressor surge is a systemic failure in which the gas flow in a compressor train reverses due to high back pressure. This is usually because of low flow rates throughout the compressor train – as a result the compressor back pressure is higher than the output pressure, resulting in the reversal of flow [22]. Vibrations occur as a result and physical damage to machinery is possible. Recovery from surge events require all compressed gas to be removed from the compressor train and the process restarted, which is time consuming and expensive [5]. Predicting and prevention of the occurrences of such events is therefore highly beneficial to process operators. Thirty surge event datasets of the compressor train illustrated in Figure 10 has been provided courtesy of Emerson Process Management and Nova Chemicals. Each dataset captures 7200 samples of data over 5 days and contains one surge event per dataset. The timing of the surge event was determined based on operator experience. It was communicated by experts at Emerson and Nova that operation four days prior to each surge event appeared to be at steady state and normal operation.

PCA was applied to the dataset, and the number of principal components retained was five due to the amount of noise present per dataset varied, meaning that consistent visualization and analysis would be difficult with varying number of principal components selected. Selecting five principal components captured 61% to 93% of the variance of the datasets. A fifth order Savitzky Golay filter was applied to the centroids prior to the construction of the confidence ellipses.

Confidence ellipses were then constructed and used as the fault detection mechanism for detecting surge. The results of the fault detection method can be found in Table 1, which provides the time required and speed of fault detection. The speed of detection (or predictive time) was calculated by finding the difference between the time of fault detection and the time of surge, which was estimated by plant operators to be at t = 5760 minutes. A positive predictive time is desirable as it shows that the method can predict the surge before it occurs or is identified. A negative predictive time would show that the method only detects the surge event after it has been identified by process operators.

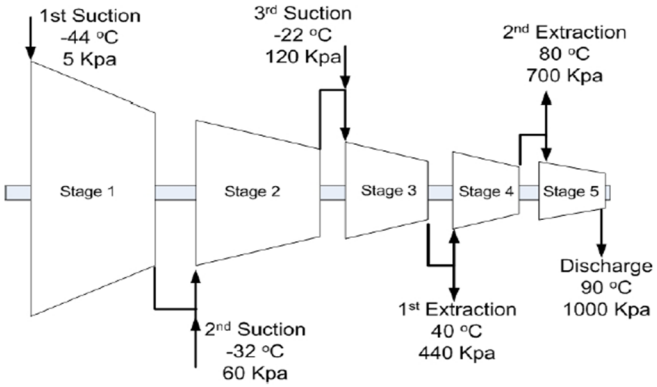


Fig 10. Schematic of compressor system [5]

Dataset #	Fault Detection Time (minutes) Operator-estimated fault time: 5760 min	Predictive time (minutes)
1	5731	29
2	6610	-850
3	4023	1737
4	4933	827
5	4596	1164
6	7110	-1350
7		
8	4099	1661
9	4159	1601
10	4244	1516
11	4974	786
12	5593	167
13	5265	495
14	5176	584
15	5810	-50
16	5700	60
17	4827	933
18	5315	445
19	5701	59
20	3502	2258
21	5448	312
22	5599	161
23	4541	1219
24	5007	753
25	5056	704
26	5867	-107
27	6388	-628
28	5151	609
29	4459	1301
30	5694	66

Table 1. Fault detection times and predictive times for 30 datasets
A negative value indicates that the fault was detected after the operator-estimated fault time
Blank cells indicate that no fault was detected by the method

Of the 30 datasets available, 29 contained detectable surge faults, and in 25 cases, the predictive times were positive (i.e., faults were detected before being declared as such by operators), suggesting that surge events may have been avoided with the proposed fault

detection scheme in place. False alarms (that is, flagged samples that cannot be related to the surge event) were found in some cases. False alarm rates for each dataset were calculated as a ratio of the number of false alarm samples over the number of samples in the provided steady state region. The average false alarm rate for the 30 datasets is 0.035.

In terms of fault classification, a comparison of the angle plots for all 30 datasets suggests that a subset of the surge events can be classified as belonging to two separate groups. Specifically, six of the 30 datasets (1, 8, 11, 15, 25, 26) belong to one group, whereas three other datasets (3, 19, 20) belong to another group.

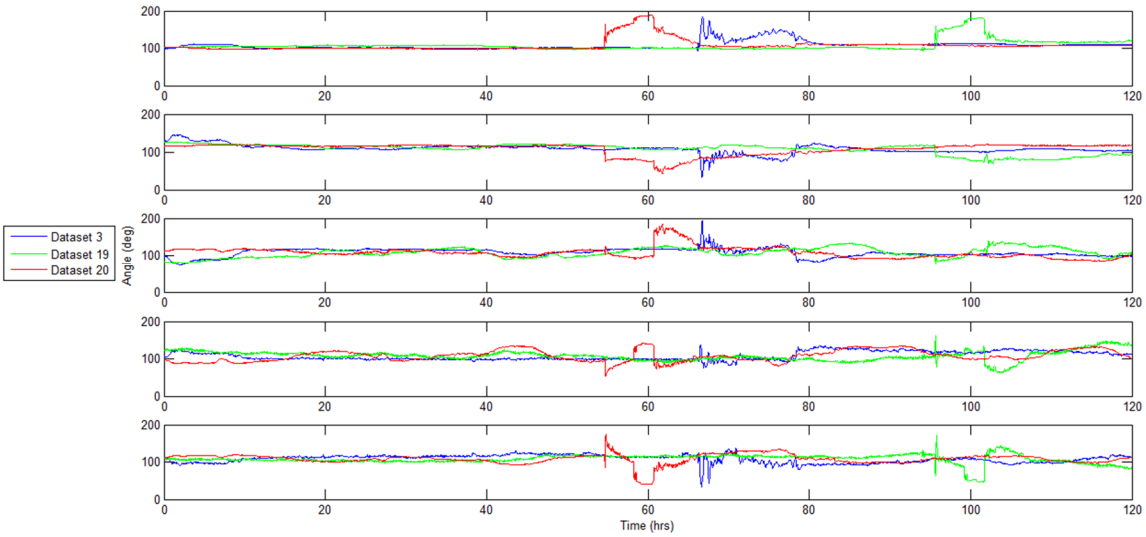


Fig 11. Plot of all 5 angles for second grouping of similar datasets

Dynamic time warping was performed on the datasets shown in Figure 11 to verify that the events are similar.

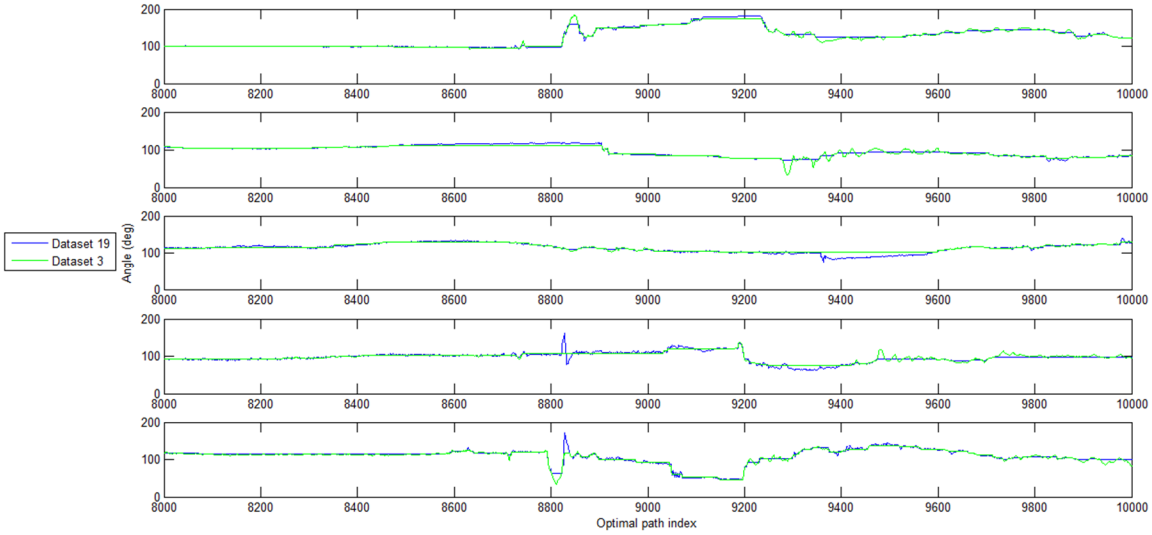


Fig 12a. Plot of interior angles across datasets 19 and 3 in Group 2 with dynamic time warping

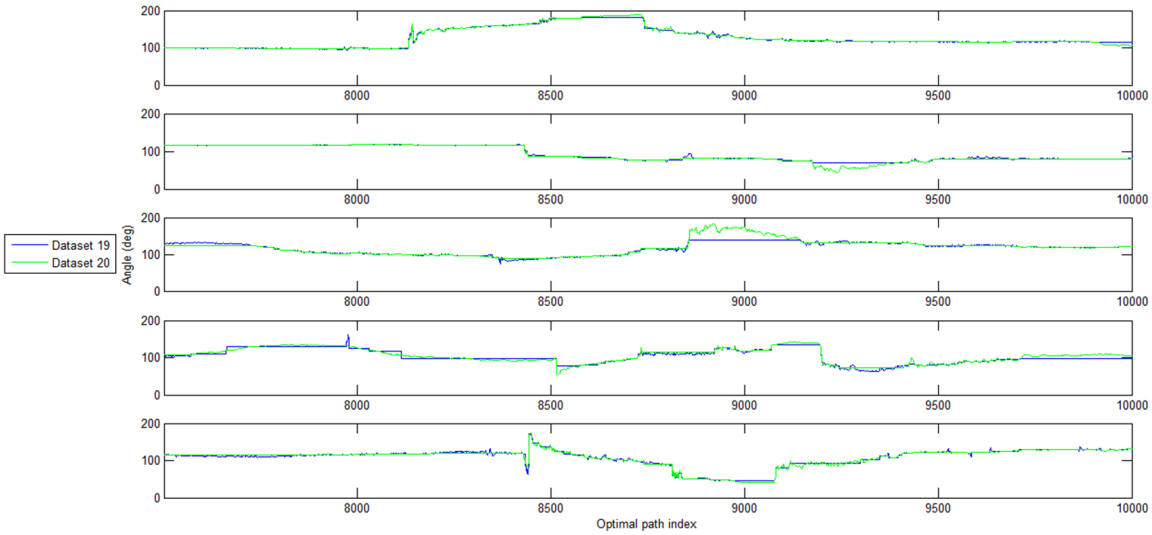


Fig 12b. Plot of interior angles across datasets 19 and 20 in Group 2 with dynamic time warping

Dataset 19 was used as the reference profile to conduct dynamic time warping on Group 2 data. Datasets 3 and 20 both matched well with the reference dataset, as seen in Figures 12a and 12b respectively. Therefore, the three datasets can be classified to be of the same fault type.

The other datasets did not fall into either of these groups due to noise, which made them difficult to categorize. Some faults appeared to be unique and did not have a good match with any of the datasets, so they could not be categorized into any group. These observations indicate that the surge events for the compressor train in question are granular – not all surge events are the same.

5.2 Column Flooding

Column flooding is a condition where upward gas flow in a distillation column prevents the liquid phase from flowing downwards, trapping the liquid in the liquid-vapor space and eventually causing the column to flood. This event is undesirable as product specifications cannot be met under such a condition. Column flooding is a systemic problem as it can happen anywhere in the column and will extend to the entire column [23].

For this case study, over four calendar months of operation data was used, with each month's data collected in its own dataset. The sample rate of the data was 1 sample per minute, for a total of about 44000 samples per dataset. Estimated times of column flooding were obtained based on operator experience. A fifth order Savitzky-Golay filter was also applied to the centroids before defining the confidence ellipses.

The confidence ellipse fault detection method was used to verify the dates of column flooding. The steady-state region used for the construction of the confidence ellipses vary per dataset and selected using the moving window variance method described in Section 3.4.2. The results are presented below.

Month (2009)	Operator -Estimated Dates of Flooding	Start Time of Flooding Event Indicated by the Proposed Method (In Days)
June	23-25	23.31
July	6-7	4.06
	11-12	9.5
	26-27	23.82
August	24	25.36
September	1-3	1
	7-8	
	16-18	13.59

Table 2. Operator-estimated dates of column flooding and the approximate dates of fault detection
Blank cells indicate that no fault was detected by the method

As seen above, all of the flooding events around the dates provided were detected. However, false alarms were also detected – they were present in the June and August datasets, which only have one detected flooding event each. This suggests that the primary source of variation was noise, which can affect the steady state region selection. The false alarm rates in the June and August dataset is 0.033 and 0.016 respectively – these false alarm rates were characterized as the ratio of the number of data samples flagged that did not belong to any of the estimated flooding dates over the total number of data samples available in the dataset. These regions may be highlighting other non-flooding events or other flooding events not noticed by operators.

Fault classification for this dataset is difficult due to process uncertainties and noise. A limited classification of the flooding events was provided by the operators. In particular, the cited reason for the flooding events in June and July was the presence of undesired temperature gradients across the column. A comparison based on angle plots was attempted between the flooding event on June 23-25 and the flooding event from July 26-27. Due to the large size of the datasets, only precursor data – data around the relevant periods – were used when performing dynamic time warping.

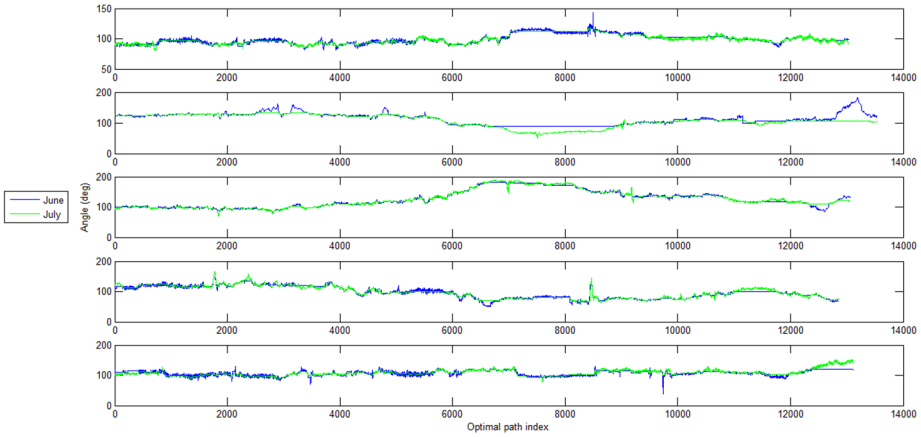


Fig 13. Plot of all 5 interior angles with dynamic time warping for the periods of 23rd to 27th of the June and July datasets

As seen above, the two flooding events appear to be of the similar type, which matches the classification provided by the operators.

Chapter 6: Benchmarking with the Tennessee Eastman Process Simulator

The Tennessee Eastman Process (TEP) was used to validate the fault detection method proposed in this report [24]. The purpose of doing so is to test the ability of the proposed mechanism to detect specific and non-systemic faults. The TEP is also a popular benchmarking tool for fault detection studies and is frequently used in literature. The work of Russell et. al., Tamura and Tsujita, as well as work by Zhang will be referred to, with each paper having done work on fault detection method(s) [12,25,26]. A simulator is available for the TEP model in MATLAB [27].

The TEP simulator has provisions for the introduction of twenty faults, five of which are not described specifically.

Fault No.	Description	Type
1	A/C feed ratio, B Composition constant (stream 4)	Step
2	B Composition, A/C ratio 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, 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	
17	Unknown	
18	Unknown	
19	Unknown	
20	Unknown	

Table 3. List of available faults in the TEP [28]

The procedure for performing the fault detection tests are as follows: For each fault, the simulator was run for at least 12 hours (720 minutes), and a fault was injected at t = 300 minutes. Fault detection using the multivariate control chart method was conducted on the data obtained from this simulator – the steady state operating region being defined based on data prior to t = 300 minutes. The number of principal components retained was five. For comparison with other methods that use T^2 and Q measurements, the approach in Dunia et.al. [5] was followed and the T^2 and Q measures obtained as a result of conducting PCA are also included as two additional axes while doing fault detection. A fifth order Savitzky-Golay filter was applied to the centroids here before calculating the confidence ellipse as well.

The results of the fault detection method are provided in Table 4 alongside the results reported in the previous papers. The detection time is the time required for a fault to be detected after it was injected.

Detection time (minutes)														
			Russell, Chiang, and Braatz [12]							Tamura and Tsujita [25]		Zhang [26]		
Fault Number	Radial Plots PCA	Radial Plots PCA w/ Q and T^2	PCA T^2	PCA Q	DPCA T^2	DPCA Q	CVA T_s^2	CVA Tr^2	CVA Q	PCA Q	PCA T^2	KPCA	KICA	Imp. KICA
1	1	1	21	9	18	15	6	9	6	21	21	15	9	6
2	19	19	51	36	48	39	39	45	75	27	33	30	36	33
3	1	1												
4	1	1		9	453	3	1386	3				9	6	6
5	1	1	48	3	6	6	3	3	0			3	3	3
6	1	1	30	3	633	3	3	3	0	15	15	3	3	0
7	1	9	3	3	3	3	3	3	0	15	15	3	3	0
8	5	22	69	60	69	63	60	60	63	54	60	75	69	60
9	73	66								894	903			
10	28	1	288	147	303	150	75	69	132	99	150	60	51	42
11	1	5	912	33	585	21	876	33	81			69	57	45
12	26	21	66	24	9	24	6	6	0			9	6	3
13	25	30	147	111	135	120	126	117	129	48	48	123	114	99
14	44	3	12	3	18	3	6	3	3			3	3	3
15	28	34		2220			2031					27	27	21
16	28	31	936	591	597	588	42	27	33	153	624	27	21	9
17	51	28	87	75	84	72	81	60	69	189	192	57	51	51
18	28	30	279	252	279	252	249	237	252	237	252	222	198	195
19	46	18				246		33						
20	16	23	261	261	267	252	246	198	216	39	45	177	165	135

Table 4. Fault detection time for Tennessee Eastman Process
Blank cells indicate that no fault was detected by the method

From Table 4, the proposed method is shown to outperform the average detection time of other methods for 18 of the 20 faults provided in the TEP simulator. Several faults were not detected by one or several methods, a shortcoming that can be attributed to the small impact the fault may have on the system states and outputs.

False detection and missed detection rates were also found for the method. A fault is only detected when two consecutive samples are flagged as a fault. A binary approach is taken – all samples prior to t = 300 minutes should not be detected as a fault, while all samples after should be considered faulty data and detected. If either condition is not met then either a false detection or a missed detection has occurred. This definition is used to calculate the false

detection and missed detection rates and is similar to the method used in Zhang to calculate detection rate [26].

Missed Detection Rates												
			Russell, Chiang, and Braatz [12]							Zhang [26]		
Fault Number	Radial Plots (PCA)	Radial Plots (PCA w/ Q and T^2)	PCA T^2	PCA Q	DPCA T^2	DPCA Q	CVA T_s^2	CVA Tr^2	CVA Q	KPCA T^2	KICA T^2	Imp. KICA T^2
1	0.0541	0.1351	0.008	0.003	0.006	0.005	0.001	0	0.003	0	0	0
2	0.7312	0.4447	0.02	0.014	0.019	0.015	0.011	0.010	0.026	0.02	0.02	0.02
3	0.9644	0.7316	0.998	0.991	0.991	0.990	0.981	0.986	0.985	0.96	0.94	0.92
4	0.5202	0.1639	0.956	0.038	0.939	0	0.688	0	0.975	0.91	0.18	0.19
5	0.7173	0.6746	0.775	0.746	0.758	0.748	0	0	0	0.75	0.71	0.71
6	0	0.0857	0.011	0	0.013	0	0	0	0	0.01	0	0
7	0.0652	0	0.085	0	0.159	0	0.386	0	0.486	0	0	0
8	0.2763	0.3246	0.034	0.024	0.028	0.025	0.021	0.016	0.486	0.03	0.03	0.02
9	0.7696	0.677	0.994	0.981	0.995	0.994	0.986	0.993	0.993	0.96	0.95	9.95
10	0.791	0.8242	0.666	0.659	0.580	0.665	0.166	0.099	0.599	0.57	0.19	0.20
11	0.7363	0.7648	0.794	0.356	0.801	0.193	0.515	0.195	0.669	0.76	0.19	0.18
12	0.4489	0.4157	0.029	0.025	0.01	0.024	0	0	0.021	0.03	0.03	0.02
13	0.3017	0.2684	0.060	0.045	0.049	0.049	0.047	0.040	0.055	0.06	0.05	0.05
14	0.8005	0.8076	0.158	0	0.061	0	0	0	0.122	0.21	0	0
15	1	0.9406	0.988	0.973	0.964	0.976	0.928	0.903	0.979	0.95	0.95	0.94
16	0.8694	0.658	0.834	0.755	0.783	0.708	0.166	0.084	0.429	0.70	0.20	0.20
17	0.6508	0.4893	0.259	0.108	0.240	0.053	0.104	0.024	0.138	0.26	0.05	0.05
18	0.5748	0.6176	0.113	0.101	0.111	0.100	0.094	0.092	0.102	0.10	0.10	0.09
19	0.9762	0.8337	0.996	0.873	0.993	0.735	0.849	0.019	0.923	0.97	0.25	0.23
20	0.0944	0.1786	0.701	0.570	0.644	0.558	0.44	0.342	0.547	0.59	0.42	0.50

Table 5. Missed detection rates for Tennessee Eastman Process
Missed detection rates not available for Tamura and Tsujita [25]

The centroid representation compares well to that of other methods.

False Detection Rates		
	Radial Plots (PCA) Radial Plots (PCA) with T^2 and Q	0.0330 0.0377
Russell, Chiang, and Braatz [12]	PCA T^2 PCA Q DPCA T^2 DPCA Q CVA T_s^2 CVA T_r^2 CVA Q	0.014 0.016 0.006 0.281 0.083 0.126 0.087
Zhang [26]	PCA T^2 KPCA T^2 KICA T^2 Improved KICA T^2	0.005 0.0152 0.0031 0.0027

Table 6. False detection rates for Tennessee Eastman Process
False detection rates not available for Tamura and Tsujita [25]

From Table 6, the false detection rates for the proposed method are similar to that of other methods in literature.

Chapter 7: Discussion and Perspective

The effectiveness of the proposed method for fault detection is shown through analysis of the case studies presented above, using both data from a process simulator and actual industrial datasets. Both localized and systemic faults are shown to be detectable using the 3D radial plot and centroid representation. The geometry of the 3D radial plots allows for further information to be extracted through the monitoring of polygonal angles in time for fault classification purposes.

For visualization purposes, the 3D radial plots method enables a time explicit multivariate representation of large datasets, which was only possible through the use of multiple score plots. The centroid representation allows us to represent the process in two dimensions while still retaining any time-sensitive characteristics of our data. Its use as a multivariate control chart is demonstrated in this report and opens up further possibilities for data analysis such as clustering methods.

Informal interactions with industrial plant personnel have revealed that the three-dimensional visualization of data are easily understood by operators and is clearly superior to parallel coordinates or multiple score plots.

However, the method, like other data-driven mechanisms for fault detection, is susceptible to noise, so pre-processing of the data to reduce the impact of outliers and noise is always desirable and will improve the results of fault detection, particularly in terms of false detection rates and missed detection rates.

A final issue with regards to the construction of confidence ellipses is that they are static and do not change in time as more data is read in. Adaptive methods for evolving the ellipse while streaming in new data is a potential direction for future work. Fault detection during process transitions between operating states may also be possible by adapting the ellipses as needed.

Chapter 8: Conclusions

In the report, a novel method for the detection of systemic process faults is presented. The method consists of creating a three-dimensional radial plot to plot multiple variables in time. The resulting visualization is a time explicit representation of multivariate data. This in turn can provide better information for operator displays compared to the currently used methods for data representation.

Faults are detected when the process samples violate the limits of a desired operating region, which is defined by a confidence ellipse that captures said operating region. It is shown that the method's performance is comparable to, if not better than, other fault detection methods in literature. The method is also used to analyze actual industrial data and is able to anticipate systemic faults in both a compressor train and a distillation column.

A mechanism for fault classification using the interior angles of the samples in the radial plot representation is also presented in the report and has the potential for real-time implementation.

The method can be applied to any dataset and a standalone program of the visualization method and fault detection mechanism may be released in the near future.

Appendix (Dataset DS1):

Observations	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7
1	52.15104	19.21382	100.9697	32.86657	64.81821	20.33766	347.1172
2	48.5197	19.97637	101.8019	32.84908	66.89273	20.66234	346.6867
3	50.98206	19.65381	102.4758	32.89453	67.28734	21.8356	346.7377
4	51.31496	17.03205	100.4126	34.62617	64.18542	22.00215	348.9064
5	49.54007	18.64558	102.7398	33.79634	64.41734	20.97363	346.8051
6	50.62845	19.00023	102.2131	34.15037	65.66244	21.68786	347.423
7	50.04591	18.84113	101.5825	33.20726	67.06497	20.70713	344.6934
8	49.10358	18.02943	101.0892	33.62383	67.34672	21.92036	345.3292
9	50.23976	19.32707	102.641	34.57129	67.53958	20.60958	345.986
10	53.72241	19.19023	101.8693	32.43254	65.64807	21.52871	346.8652
11	51.65718	20.62504	101.9434	32.9471	65.36202	21.59507	347.0304
12	52.68589	19.4452	104.0915	34.71172	64.96093	22.64062	349.1286
13	50.93906	20.02596	101.9407	35.37108	66.52062	20.06329	344.8115
14	51.14075	19.17759	102.5501	32.79897	66.70951	21.35153	344.7907
15	50.82536	17.47007	101.3105	34.08699	64.38965	19.37649	364.7247
16	48.71859	17.80987	103.151	32.38646	65.7552	21.19504	364.0869
17	49.21272	18.51945	101.2414	34.27679	65.33914	21.81811	363.8686
18	49.25756	18.82497	100.3293	33.73811	66.44828	19.93257	365.8114
19	49.71707	19.82305	101.2957	32.57727	66.96602	20.76101	364.3838
20	100.8204	52.15818	67.13404	17.50275	89.74815	28.27168	407.6089
21	101.959	52.35225	70.09609	18.24484	89.01294	28.6664	407.2975
22	103.7175	53.22556	67.64299	16.80958	88.08405	26.9809	405.448
23	103.7005	52.39659	69.09773	16.05571	88.00298	28.43369	407.7703
24	104.1793	51.71946	69.71967	15.82389	87.28477	26.75917	406.7051
25	105.6697	52.3834	69.27764	15.68788	89.60259	26.01458	406.2575
26	50.16743	18.48934	101.6088	33.9001	64.76961	20.64649	345.8122
27	50.59506	19.85109	100.3611	33.8298	63.70124	18.98059	346.32
28	51.76132	19.37258	101.1612	32.69606	64.803	19.87094	346.9945
29	50.86219	19.68772	102.1378	34.22339	67.06336	20.05118	345.9562
30	50.30873	18.67364	100.4354	32.39403	65.84833	22.0273	344.0958
31	51.26709	21.24216	99.95538	34.93778	65.21288	20.84922	346.6231
32	52.41266	19.22711	100.485	31.31406	65.76906	22.64998	345.2125
33	49.66752	18.19126	102.0312	33.15998	65.93934	21.03158	347.5765
34	51.03453	17.78108	100.8957	34.17964	65.97606	20.92217	346.3157
35	49.37383	19.19749	101.3873	33.43182	65.34612	20.32841	346.3364
36	48.941	18.6869	99.41433	33.46247	67.80024	19.64577	345.3201
37	49.51367	17.73693	100.3818	32.8994	66.5646	20.69357	344.799
38	50.11895	18.61487	100.4508	35.36838	63.02115	20.13637	344.8958
39	48.23254	17.45313	99.88083	32.93868	64.5396	18.94534	344.3149
40	48.59724	19.76062	102.3178	34.0936	65.45264	21.0093	344.7761

References

[1] M. Kano, K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss, Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem, Comput. Chem. Eng. 26 (2002) 161–174.

[2] S. Yoon, J.F. MacGregor, Fault diagnosis with multivariate statistical models part I: Using steady state fault signatures, J. Process Control. 11 (2001) 387–400.

[3] V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri, K. Yin, A review of process fault detection and diagnosis: Part III: Process history based methods, Comput. Chem. Eng. 27 (2003) 324–346.

[4] A. Inselberg, The plane with parallel coordinates, Vis. Comput. 1 (1985) 69–91.

[5] R. Dunia, T.F. Edgar, M. Nixon, Process monitoring using principal components in parallel coordinates, AIChE J. 59 (2013) 445–456.

[6] R. Dunia, G. Rochelle, T.F. Edgar, M. Nixon, Multivariate Monitoring of a Carbon Dioxide Removal Process, Comput. Chem. Eng. (n.d.).

[7] Jamal Alsakran, Ye Zhao, Xinlei Zhao, Tile-based parallel coordinates and its application in financial visualization, in: Vis. Data Anal. 2010, San Jose, California, 2010.

[8] M. Novotny, Visually Effective Information Visualization of Large Data, in: Proc. 8th Cent. Eur. Semin. Comput. Graph., 2004: pp. 41–48.

[9] M. Novotny, H. Hauser, Outlier-Preserving Focus+Context Visualization in Parallel Coordinates, Vis. Comput. Graph. IEEE Trans. On. 12 (2006) 893–900.

[10] C. Tominski, J. Abello, H. Schumann, Interactive poster: 3D axes-based visualizations for time series data, in: Poster Compend. IEEE Symp Inf. Vis. InfoVis 05, Minneapolis, USA, 2005.

[11] D. Said, Radar Plot, n.d.

[12] E.L. Russell, L.H. Chiang, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemom. Intell. Lab. Syst. 51 (2000) 81–93.

[13] B.M. Wise, N.B. Gallagher, S.W. Butler, D.D. White, G.G. Barna, A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, J. Chemom. 13 (1999) 379–396.

[14] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemom. Intell. Lab. Syst. 28 (1995) 3–21.

[15] J.F. MacGregor, T. Kourti, Statstical process control of multivariate processes, Control Eng. Pract. 3 (1995) 403–414.

[16] P. Nomikos, J.F. MacGregor, Monitoring batch processes using multiway principal component analysis, AIChE J. 40 (1994) 1361–1375.

[17] V. Venkatasubramanian, R. Rengaswamy, K. Yin, S.N. Kavuri, A review of process fault detection and diagnosis: Part I: Quantitative model-based methods, Comput. Chem. Eng. 27 (2003) 293–311.

[18] V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri, A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies, Comput. Chem. Eng. 27 (2003) 313–326.

[19] S.J. Qin, Survey on data-driven industrial process monitoring and diagnosis, Annu. Rev. Control. 36 (2012) 220–234.

[20] M. Kim, S.H. Yoon, P.A. Domanski, W.V. Payne, Design of a steady-state detector for fault detection and diagnosis of a residential air conditioner, Int. J. Refrig. 31 (2008) 790–799.

[21] Y. Zhang, T.F. Edgar, A Robust Dynamic Time Warping Algorithm for Batch Trajectory Synchronization, in: Seattle, Washington, 2008.

[22] J. Cahill, Surge Control Considerations in Centrifugal Compressors, Emerson Process Experts. (2010).

[23] Distillation Column Flooding Diagnostics with Intelligent Differential Pressure Transmitter, (n.d.).

[24] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, Comput. Chem. Eng. 17 (1993) 245–255.

[25] M. Tamura, S. Tsujita, A study on the number of principal components and sensitivity of fault detection using PCA, Comput. Chem. Eng. 31 (2007) 1035–1046.

[26] Y. Zhang, Fault Detection and Diagnosis of Nonlinear Processes Using Improved Kernel Independent Component Analysis (KICA) and Support Vector Machine (SVM), Ind Eng Chem Res. 47 (2008) 6961–6971.

[27] N.L. Ricker, Tennessee Eastman Challenge Archive, (n.d.).

[28] J.-M. Lee, C. Yoo, I.-B. Lee, Statistical monitoring of dynamic processes based on dynamic independent component analysis, Chem. Eng. Sci. 59 (2004) 2995–3006.