Proceedings of the American Control Conference
Arlington, VA June 25-27, 2001

Matching Patterns from Historical Data Using PCA and Distance Similarity Factors

Ashish Singhal[†] Dale E. Seborg[‡]

Department of Chemical Engineering
University of California, Santa Barbara, CA 93106

Abstract

The diagnosis of abnormal plant operation can be greatly facilitated if periods of similar plant performance can be located in the historical database. A novel methodology is proposed for this pattern matching problem. The new approach provides a preliminary screening of large amounts of historical data in order to generate a candidate pool of similar periods of operation. This much smaller number of records can then be further evaluated by someone familiar with the process. Similarity factors are used to characterize the degree of similarity between the current abnormal operation and historical data. A new Distance Similarity Factor is proposed that complements the standard PCA similarity factor. The two similarity factors provide the basis for an unsupervised pattern matching technique.The proposed pattern matching methodology has been evaluated in a detailed case study for a controlled CSTR (14 measured variables, more than 474,000 data points for each measured variable, and 19 operating modes/faults). The proposed methodology was able to locate over 90% of the previous occurrences of "abnormal situations".

1 Introduction

Advances in data collection technology have resulted in routine collection and storage of large volumes of data in industrial plants. Large plants record thousands of process variables, product quality, production, and maintenance information on a frequent basis. Thus, massive amounts of stored data can be used for analysis of the process and previous occurrences of abnormal situations. A historical database contains potentially valuable process information, but it is notoriously difficult to extract it. Industrial plants have therefore been called, "data rich, but information poor". The problem of extracting valuable information from large historical database has received considerable attention in other fields, as indicated by the growing interest in data mining and knowledge discovery problems (Apté, 1997; Agrawal et al., 1998). An industrial consortium has estimated that abnormal, but preventable, plant behavior costs the U.S. petrochemical industry more than $20 billion per year (Honeywell, Inc., 2000). This estimate is just one indication that improved plant monitoring can play a key role in increasing plant productivity, an important concern in the intensely competitive global economy.

1.1 Abnormal situation analysis

In this paper, an abnormal situation is defined as an unanticipated plant situation that has (or could have) serious consequences, but does not warrant drastic action such as an emergency shutdown. For example: (i) several key control loops exhibit unusual oscillatory behavior for several hours before the oscillations die out, and (ii) the product quality is low for a period of several days, for no apparent reason. After an abnormal situation is detected, it is important to diagnose its cause in order to take corrective action, and to prevent future occurrences.

A wide variety of fault diagnosis techniques are available in the process monitoring literature (Frank, 1990; Kramer and Mah, 1994; Kourti and MacGregor, 1996). However, existing techniques rely on one or more of the following types of information: current plant data, previous experience, or process knowledge that is codified in the form of a process model or a knowledge-based system. A valuable resource, historical plant data, is seldom considered in a systematic manner during fault diagnosis.

If the same type of abnormal situation has occurred in the past, the relevant historical data would be a valuable source of information for difficult process diagnosis problems. This additional information can facilitate two important activities: (i) identifying the root cause of the abnormal situation, and (ii) developing an effective remedy that will prevent future occurrences or minimize their impact. These considerations motivate the main premise of this paper:

> *After an abnormal plant situation occurs, it would be very beneficial to be able to efficiently search a historical database in order to locate periods of similar, but not necessarily identical, plant behavior.*

Emphasis is placed on performing a preliminary screening of historical data because an efficient screening technique can be used to narrow the search for similar periods of process behavior by identifying a relatively small number of promising data records within the historical database. These records will be referred to as the *candidate pool*. Then a person familiar with the process (a process expert) could take a closer look at the candidate pool in order to narrow the search further and to diagnose the root cause of the abnormal situation.

Data mining of historical databases has received attention in the computer science literature; however problems involving time-series databases have been addressed only recently. Wang and McGreavy (1998) used clustering methods to classify abnormal behavior of a refinery fluid catalytic cracking process. Ng and Huang (1999) also used a clustering approach to classify different types of stars based on their light curves. In a previous study, Johannesmeyer and Seborg (1999) developed an efficient technique to locating similar records in the historical database using PCA similarity factors. This paper introduces a new *Distance Similarity Factor* to characterize the distance between the subspaces spanned by two datasets. The standard PCA similarity factor and the new distance similarity factor are used to generate the candidate pool.

[†]E-mail: ashishs@engineering.ucsb.edu
[‡]E-mail: seborg@engineering.ucsb.edu, Corresponding author

2 Methodology

The proposed pattern matching strategy is summarized in the flowchart in Figure 1. First, the user defines the *snapshot* of the data that serves as a template for searching the historical database. The snapshot specifications consist of: (i) the relevant variables, and (ii) "duration of the abnormal situation". These well defined specifications can be arbitrarily chosen; no special plant tests or pre-imposed conditions are necessary.

In order to compare the snapshot data to historical data, the relevant historical data are divided into data windows that are the same size as the snapshot data. The historical data sets are then organized by placing windows side-by-side along the time axis, which results in equal length, non-overlapping segments of data. The similarity between these windows of historical data and the current data can then be calculated via appropriate similarity measures.

Once the historical data has been divided into data windows, the snapshot data is compared to these data windows using appropriate similarity measures. The similarity measure is a number between zero and one, where zero denotes no similarity and one denotes identical data sets. In this paper, similarity factors based on PCA and distance between the two datasets are developed. These similarity factors are used to define similarity between the snapshot and historical datasets. A cutoff value for the similarity factors is used so that data sets that have similarity factor greater than or equal to the cutoff are labeled as "similar". The historical data sets that are "similar" to the snapshot data are collected in a *candidate pool*. The records in the candidate pool are then given to a process expert for a detailed evaluation.

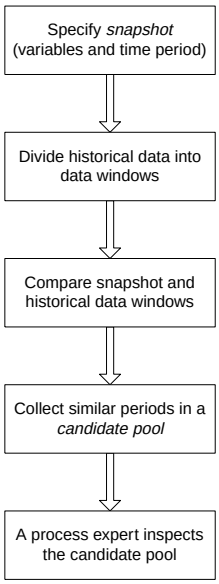

Figure 1. Proposed pattern matching approach.

2.1 PCA Similarity Factor

Because principal component analysis (PCA) has been widely reported in the process monitoring literature (Kourti and MacGregor, 1996; Martin and Morris, 1996), only a brief summary will be presented here. PCA is a multivariate statistical technique which calculates the principal directions of variability in data, and transforms the original set of correlated variables into a new set of uncorrelated variables. The new uncorrelated variables are linear combinations of the original variables. These principal components represent the most important directions of variability in a dataset (Jackson, 1991; Jolliffe, 1986).

Krzanowski (1979) developed a method for measuring the similarity of two data sets using a PCA similarity factor, S_{PCA}. Consider two data sets which contain the same n variables but not necessarily the same number of measurements. We assume that the PCA model for each data set contains k principal components, where $k \leqslant n$. The number of principal components (PC) is chosen such that k PCs describe at least 95% of the total variance in each dataset. The similarity between the two data sets is then quantified by comparing their principal components. The appeal of the similarity factor approach is that the similarity between two data sets is quantified by a single number, S_{PCA}.

Consider a current snapshot data set S and a historical data set H having the same n variables. Let the PCA models for S and H consist of k PC's each. The corresponding $(n \times k)$ subspaces are denoted by L and M respectively. The S_{PCA} compares these subspaces and is defined to be (Krzanowski, 1979),

$$S_{PCA} = \frac{trace\left(L^T M M^T L\right)}{k} \tag{1}$$

The geometric interpretation of S_{PCA} is that it is the sum of the squares of the cosines of the angles between each principal component of L and M. Thus,

$$S_{PCA} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} \cos^2 \theta_{ij} \tag{2}$$

Because subspaces L and M contain the k most important principal components that account for most of the variance in their corresponding data sets, S_{PCA} is also a measure of similarity between the data sets S and H.

2.2 Distance Similarity Factor

In this section, a distance similarity factor, S_{dist}, is introduced that compares two datasets that have the same spatial orientation but are located far apart. The new similarity factor is particularly useful when two data windows have similar principal components but the numerical values of the process variables are very different. The distance similarity factor can be used to distinguish between these two cases.

The Mahalanobis distance, Φ, from the center of the historical dataset ($\bar{x}_H$) to the center of the current snapshot dataset, $\bar{x}_S$, is defined as,

$$\Phi = \sqrt{(\bar{x}_H - \bar{x}_S)^T \Sigma_S^{*-1} (\bar{x}_H - \bar{x}_S)} \tag{3}$$

where $\bar{x}_S$ and $\bar{x}_H$ are sample mean vectors. The distance similarity factor is proposed as the probability that the center of the historical dataset, $\bar{x}_H$, is not closer than its Mahalanobis distance, Φ:

$$S_{dist} \triangleq \sqrt{\frac{2}{\pi}} \int_{\Phi}^{\infty} e^{-z^2/2} dz \tag{4}$$

Matrix Σ_S^{*-1} is the pseudo-inverse of Σ_S and is calculated using a singular value decomposition. The error function in Eq. (4) can be evaluated using standard tables or software packages. The distance similarity factor provides a natural complement to the PCA similarity factor. In contrast, the use of alarm limit violations information proposed by Johannesmeyer and Seborg (1999), require that alarm violations actually occur. This places a restriction on the analysis. The distance similarity factor proposed in this paper is independent of the alarm limits and does not use any *a priori* information.

2.3 Measures of Effectiveness of Search Techniques

Johannesmeyer and Seborg (1999) developed two useful metrics to describe the effectiveness of pattern matching techniques. They are based on the following statistics for the candidate pool and the historical database:

N_P: The size of the candidate pool. N_P is the total number of historical data records that have been labeled "similar" by the pattern matching technique.

N_1: The number of records in the candidate pool, that are actually similar to the snapshot dataset, i.e., the correctly identified records.

N_2: The number of records in the candidate pool, that are not similar to the current snapshot, i.e., incorrectly identified records. By definition, $N_1 + N_2 = N_P$.

N_{DB}: The number of historical records similar to the current snapshot data set S. Note that N_{DB} is independent of the size of the candidate pool.

Based on the above quantities, the pool accuracy p, and the search efficiency η, are defined as follows (Johannesmeyer and Seborg, 1999):

$$p \triangleq \frac{N_1}{N_P} \times 100\% \tag{5}$$

$$\eta \triangleq \frac{N_1}{N_{DB}} \times 100\% \tag{6}$$

An effective search technique should produce a high pool accuracy, as well as a high search efficiency. It is convenient to use an average of p and η as a measure of the overall effectiveness.

$$\text{Average} \triangleq \frac{p + \eta}{2} \tag{7}$$

A higher average value means better pattern matching.

3 Simulation Example: Continuous Stirred Tank Reactor

An extensive simulation case study was used to evaluate the performance of alternative pattern matching techniques for a wide variety of operating conditions and fault scenarios.

A non-isothermal continuous stirred tank reactor (CSTR) with cooling jacket dynamics and variable liquid level was simulated in order to generate historical data. A first order irreversible reaction, A → B, is assumed. A schematic diagram of the CSTR and feedback control system is shown in Figure 2. A dynamic model for the CSTR can be derived based on the assumptions of perfect mixing and constant physical parameters (Russo and Bequette, 1996). The nominal operating conditions, control structure and the controller parameters are described in detail by Johannesmeyer (1999). The historical database for the CSTR case study was designed to include both normal operating periods and a wide variety of abnormal situations or "faults".

Many fault detection and diagnosis studies have been conducted using CSTR models (Sorsa and Koivo, 1993; Vaidyanathan and Venkatasubramanian, 1992), and a large number of possible fault conditions can be considered. The 19 operating conditions in Table 1 were chosen in order to simulate the wide range of disturbance and fault types that can be encountered in a typical historical database. Fault conditions included process disturbances (e.g., ramp change in T_{CF}, step or sinusoidal changes in Q_F, etc.), instrumentation faults (e.g., dead coolant flow measurement, bias in reactor temperature measurement, etc.) and equipment faults (e.g., valve stiction, heat exchanger fouling, catalyst deactivation, etc.). Nominal sizes for the faults were chosen so that the key process variables were affected by approximately the same magnitude for each fault. Setpoint changes in reactor temperature were also included. Gaussian noise was also added to all measurements (Johannesmeyer, 1999).

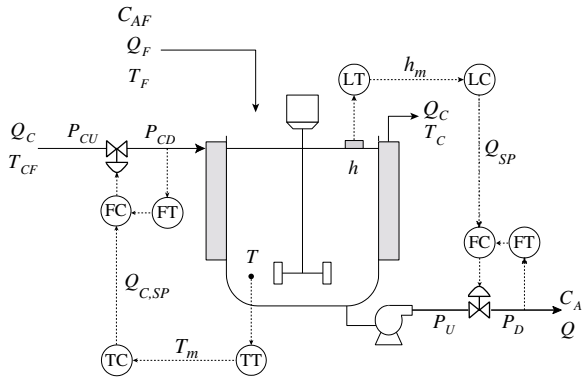

Figure 2. Schematic of CSTR system with cascade control.

3.1 Generation of Historical Database

The database was generated by simulating the controlled process via SIMULINK™in MATLAB™5.3 on an HP 715/100 UNIX workstation. In order to generate a large historical database, the CSTR system was simulated for a period of 39 days with measurements being recorded every five seconds. Measurements of the 14 process variables given in Table 2 were included in the database. The last four measurements correspond to controller output signals. For example, hC is the output signal in mA for the level controller.

The historical database was generated in the following manner. Each period of operation lasted 120 minutes. Each consecutive mode of operation (i.e., fault type, set point change, or normal operation) to be simulated was chosen randomly from the list in Table 1. The fault direction and magnitude were also randomly selected for each period of operation. The fault direction could be positive or negative for faults that contain ramps or steps. The fault magnitude was chosen randomly to be between 25% to 125% of the nominal fault magnitude. Once the mode of operation and any necessary parameters (i.e., direction and magnitude) were selected, the simulation ran for 120 minutes before the next period of operation began. Each 120 minute period of operation consisted of 85.3 minutes for the "event", followed by a period of 35 minutes for the process to return to the original steady state before the next period of operation began. Thus, the "event" data consisted of over 474,000 data points for each measured variable. Also, the faults occurred one at a time (i.e., no simultaneous faults) and for the same duration. The simulation generated approximately 39 days of data and 463 periods of operation, with each period containing 1024 data points for each of the 14 variables.

4 Results and Discussion

In this section, we compare the performance of alternative pattern matching techniques for the CSTR case study. For the PCA model development, the historical data were partitioned into data windows (H_i) that were the same size as the snapshot data (S). Each H_i was scaled to zero mean and unit variance. When the current snapshot, S, was compared to a data window, H_i, it was scaled using the scaling factors for H_i. Alternative pattern matching methods were compared in terms of the pool accuracy (p) and search efficiency (η), that were defined in Eqs. (5) and (6). The results are reported as

average values for the 19 operating modes, where each mode was designated in turn as the "abnormal" situation for the snapshot data.

4.1 Results for the PCA Similarity Factor

In the proposed methodology, two data sets S and H are considered similar if the similarity factor exceeds a specified threshold or *cutoff value*. The effect of the cutoff value on the performance of the PCA similarity factor, S_{PCA}, is illustrated in Figure 3. The values of p, η and their average are the mean values for the 19 operating conditions in Table 1. Thus the nominal condition for each operating mode was considered in turn to be the "snapshot data", and each data window H_i in the historical database was screened. As the cutoff value increases, the proportion of correctly identified records in the candidate pool, p, increases, but the total number of correctly identified records decreases because η decreases. This is reasonable because increasing the cutoff results in only highly similar records appearing in the candidate pool; this increases p, but decreases η. But the average value is relatively constant for most part of the indicated range and has a maximum value of 70% at a cutoff value of 0.965.

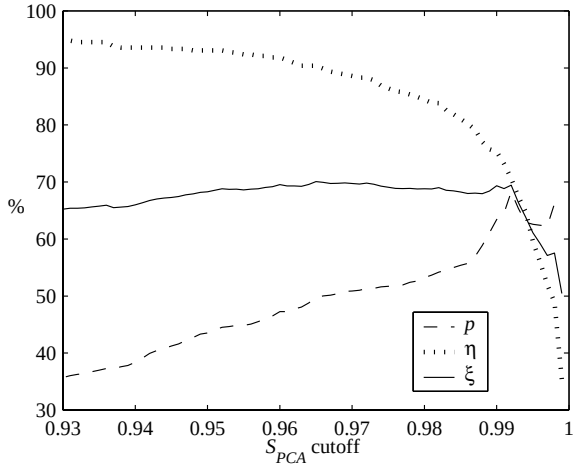

Figure 3. Effect of PCA similarity factor cutoff on pattern matching.

4.2 Results Using Both the PCA and Distance Similarity Factors

Additional information about the distance between the subspaces of the S and H datasets can provide further insight into their similarity or dissimilarity. In particular, the new distance similarity factor, S_{dist}, can be used in conjunction with S_{PCA}, to provide further refinement of the results.

When both S_{PCA} and S_{dist} are employed in the analysis, the historical data set H_i is considered to be similar to the snapshot data set S, if,

$$S_{PCA} \geqslant \theta_{PCA} \quad \text{and} \quad S_{dist} \geqslant \theta_{dist} \tag{8}$$

where θ_{PCA} and θ_{dist} are the cutoff values for the PCA and distance similarity factors, respectively. To find the best performance when both similarity factors are used, the average values for the 19 operating modes is plotted as a function of the two cutoffs as a 3-D surface in Figure 4. The best performance was achieved for θ_{PCA} = 0.965 and θ_{dist} = 0.290. A section of this surface cut at the optimum value of the distance similarity factor is shown in Figure 5. A comparison of Figures 3 and 5 indicates that significantly better performance is obtained when both the PCA and distance similarity factors are used.

A comparison of the performance of different pattern matching techniques is presented in Table 3. The first three rows contain a summary of results obtained by matching the T^2, Q and the combined discriminant (Raich and Çinar, 1994) statistics of the historical data record H_i with those for the snapshot data S. It can be seen that these statistics do not produce very satisfactory results. A comparison of the last two rows in Table 3 indicates that the addition of the distance similarity factor produces a dramatic improvement in the p and average values (40% and 17%), while only reducing η by a small amount. The size of the candidate pool (N_P), also serves as a diagnostic measure of how well a technique performs by comparing it with the average size of the candidate pool if it only contained correctly identified records (i.e., if p = 100%). This average size is 17. Use of the distance similarity factor with the PCA similarity factor reduces the candidate pool size from 34 to 17 which is exactly equal to the average number of similar records in the historical database. Therefore, the combination of the PCA and distance similarity factor produces very accurate results.

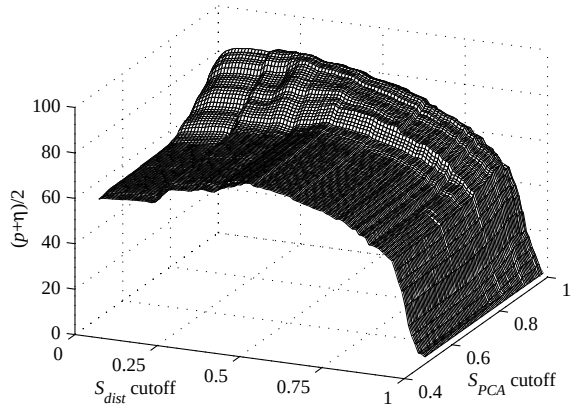

Figure 4. Effect of PCA and distance similarity factor cutoffs.

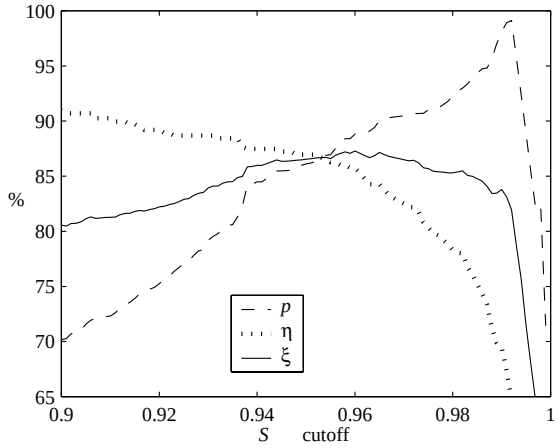

Figure 5. Effect of PCA similarity factor cutoff on pattern matching. θ_{dist} = 0.290.

It may be noted that the new pattern matching approach can be applied without specifying cutoff values. Instead, the person familiar with the process could first evaluate the historical data windows which have the largest S_{PCA} and S_{dist} values. The evaluation could stop when an appropriate number of successful pattern matches have been confirmed or when a desired value of N_P has

been reached.

The computational requirements for calculation of the similarity factors are modest. For example, it takes less than 10 seconds on a Pentium III/550 MHz computer running MATLAB™version 5.3 to build PCA models on the current snapshot and all the 463 historical data sets (1024 data points per variable per data set), and to calculate both the similarity factors for all historical data sets.

5 Conclusions

A novel methodology has been developed for locating similar periods of historical data, similar to an "abnormal situation", that is of interest. The proposed pattern matching methodology is both data driven (does not use process models or prior process knowledge), and unsupervised. The new approach is based on principal component analysis and a new metric for the distance between the two datasets. The computational load is modest, which allows processing of large amounts of process data in relatively short time.

In an extensive simulation case study, the proposed approach performed better than the existing PCA methodology for a wide range of operating conditions and faults. The combination of PCA and distance similarity factors provide an effective way of matching patterns in multivariate time-series datasets.

Acknowledgements

The authors thank the UCSB Process Control Consortium and Pavilion Technologies, Inc., Austin, Texas, for providing financial support for this research.

Literature Cited

Agrawal, R., P. Stoloroz and G. Piatetsky-Shapiro (eds.). *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA (1998).

Apté, C. Data Mining: An Industrial Research Perspective. *IEEE Trans. Computational Sci. Eng.*, **4**(2), 6–9 (1997).

Frank, P. M. Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge Based Redundancy - A Survey and Some New Results. *Automatica*, **26**, 459–474 (1990).

Honeywell, Inc. Honeywell Abnormal Situation Management: Joint Research and Develepment Consortium Web Page (http://www.iac.honeywell.com/Pub/AbSitMang/) (2000).

Jackson, J. E. *A User's Guide to Principal Components*. John Wiley, NY (1991).

Johannesmeyer, M. C. *Abnormal Situation Analysis Using Pattern Recognition Techniques and Historical Data*. M.Sc. Thesis, University of California, Santa Barbara, CA (1999).

Johannesmeyer, M. C. and D. E. Seborg. Abnormal Situation Analysis Using Pattern Recognition Techniques. *AIChE Annual Meeting, Dallas, TX* (1999).

Jolliffe, I. T. *Principal Component Analysis*. Springer-Verlag, NY (1986).

Kourti, T. and J. F. MacGregor. Multivariate SPC Methods for Process and Product Monitoring. *J. Quality Tech.*, **28**, 409–428 (1996).

Kramer, M. A. and R. S. H. Mah. Model Based Monitoring. In *Proc. 2nd Int. Conf. on Foundations of Computer Aided Process Operations*, 45–71. CACHE, Austin, TX (1994).

Krzanowski, W. J. Between-Groups Comparison of Principal Components. *J. Amer. Stat. Assoc.*, **74**(367), 703–707 (1979).

Martin, E. B. and A. J. Morris. An Overview of Multivariate Statistical Process Control in Continuous and Batch Performance Monitoring. *Trans. Inst. Meas. and Control*, **18**, 51–60 (1996).

Ng, M. K. and Z. Huang. Data-Mining Massive Time-Series Astronomical Data: Challenges and Solutions. *Information and Software Tech.*, **41**, 545–556 (1999).

Raich, A. C. and A. Çinar. Statistical Process Monitoring and Disturbance Isolation in Multivariate Continuous Processes. In *Proc. IFAC-ADCHEM'94*. Kyoto, Japan (1994).

Russo, L. P. and B. W. Bequette. Effect of Process Design on the Open-Loop Behavior of a Jacketed Exothermic CSTR. *Comput. Chem. Eng.*, **20**, 417–426 (1996).

Sorsa, T. and H. Koivo. Application of Artificial Neural Networks in Process Fault Diagnosis. *Automatica*, **29**, 843–849 (1993).

Vaidyanathan, R. and V. Venkatasubramanian. Representing and Diagnosing Dynamic Process Data Using Neural Networks. *Eng. Appl. of Artificial Intelligence*, **5**, 11–21 (1992).

Wang, X. Z. and C. McGreavy. Automatic Classification for Mining Process Operational Data. *Ind. Eng. Chem. Res.*, **37**, 2215–2222 (1998).

Table 1. Modes of Operation.

ID	Operating Condition
Normal	Normal operation
F1	Catalyst deactivation by ramp increase in activation energy
F2	Heat exchanger fouling leading to ram decrease in heat transfer coefficient
F3	Dead coolant flow measurement
F4.	Bias in reactor temperature measurement
F5	Coolant valve stiction
F6	Step change in feed flow rate, Q_F
F7	Ramp change in feed concentration, C_{AF}
F8	Ramp change in feed temperature, T_F
F9	Ramp change in coolant feed temperature, T_{CF}
F10	Step change in upstream pressure in the cooling line
F11	Step change in downstream pressure in the reactor outlet line
F12	Damped oscillations in feed flow rate
F13	Autoregressive disturbance in feed flow rate
S1	Set point change in reactor temperature, T
O1	High frequency oscillations of 3 cycles/min in feed flow rate
O2	Intermediate frequency oscillations of 1 cycles/min in feed flow rate
O3	Intermediate frequency oscillations of 0.5 cycles/min in feed flow rate
O4	Low frequency oscillations of 0.2 cycles/min

Table 2. Measurements for the CSTR simulation.

C_A	T	T_C	h	Q	Q_C	Q_F
C_{AF}	T_F	T_{CF}	hC	QC	TC	Q_CC

Table 3. Best performance of the similarity factors.

Method	Best Cutoff(s)	N_P	p (%)	η (%)	Average (%)
T^2 statistic	N/A	39	25	33	**29**
Q statistic	N/A	30	61	53	**57**
Combined Q and T^2	N/A	34	66	65	**65**
PCA similarity factor alone	0.965	34	50	90	**70**
Distance similarity factor alone	0.290	138	29	94	**61**
PCA and Distance similarity factors	**0.965, 0.290**	**17**	**90**	**84**	**87**