Chemometrics and Intelligent Laboratory Systems 148 (2015) 20–31

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

An extensive reference dataset for fault detection and identification in batch processes

Jan Van Impe, Geert Gins *

KU Leuven, Department of Chemical Engineering, Chemical & Biochemical Process Technology & Control (BioTeC +), Gebroeders De Smetstraat 1, 9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form 25 August 2015
Accepted 26 August 2015
Available online 3 September 2015

Keywords:
Batch processes
Statistical process control (SPM)
Fault detection
Fault identification
Benchmark dataset

ABSTRACT

Close process monitoring (i.e., detection and identification of disturbances) is important to achieve high process efficiency and safety. The Tennessee Eastman process is an extensive benchmark dataset for fault detection and identification, but it is only representative for continuous processes because it does not contain the inherent non-stationarity that complicates monitoring of batch processes. Nevertheless, batch processes also play an important role in many types of industry. This paper therefore presents an extensive reference dataset for benchmarking data-driven methodologies for fault detection and identification in batch processes.

The original Pensim model [10] is expanded with sensor noise. By changing the properties of the initial conditions and/or model parameters, four subsets of different complexity are generated, each containing 400 batches with normal operation. To correctly assess the fault detection and identification in batch processes, 15 faults are simulated with various amplitudes and onset times for a total of 22,200 faulty batches for each subset, or 90,400 batches in total.

Analysis of the data indicates that the presented types of process faults and their various amplitudes in each of the four subsets present a suitable benchmark for fault detection and identification in batch processes. The dataset is freely available at http://cit.kuleuven.be/biotec/batchbenchmark.

1. Introduction

Modern process industry sees a major push towards safe, sustainable, and more profitable operation. Timely detection and diagnosis of process faults, before they have the opportunity to influence process safety and/or product quality, are of utmost importance to maintain safe operation and reduce or even avoid productivity losses [81,63,33]. Therefore, considerable research attention has been paid to the area of process monitoring (also called Fault Detection and Identification/isolation; FDI) over the last few decades [63,33,25].

The existing process monitoring approaches can be categorized as either model-based or data-driven [94,33].

A model-based monitoring scheme employs available first-principles models of the process under study (such as laws of motion, mass balances, energy balances, known reaction schemes, …) to detect deviations from normal operation. One of the drawbacks of model-based process monitoring is that it is limited to well-known systems of limited size [94]. Typically, first-principles models are available for mechanical or electrical systems. Chemical, biochemical, steel, pulp and paper, or semiconductor processes contain too much uncertainty (e.g., imperfect mixing, biological variability, …) or are of a too large scale to build accurate-enough first-principles models in an acceptable time [94,81,91,33].

Data-driven process monitoring, on the other hand, uses only available process measurements to characterize the nominal process operation. Next, Statistical Process Monitoring (SPM) is used to detect deviations from this normal situation. A detailed overview of active research directions and successful applications of SPM can be found in, i.a., Venkatasubramanian et al. [80], Kourti [46,47], Hwang and Kim [42], Bogomolov [12], MacGregor and Cinar [57], Qin [63], Aldrich and Auret [5], Ge et al. [33], and Ding [25].

SPM algorithms were originally developed for continuous processes because these processes operate around a steady state regime. Batch processes, on the other hand, present a much greater challenge for monitoring owing to their inherent non-stationarity, finite duration, non-linear response, and batch-to-batch variability [23,71,28]. Furthermore, batch processes commonly suffer from a lack of suitable in-line instrumentation in practice [23]. As a result, most novel techniques for fault detection and identification are still developed almost exclusively for continuous processes. Nevertheless, batch processes are widely used in a broad range of sectors, such as the chemical, pharmaceutical, or life sciences industries [28]. Therefore, the development of proper monitoring tools for batch processes is important [80]. In their review of SPM for batch processes, Yao and Gao [91] and Qin [63] reach the conclusion that more research is needed before advanced

* Corresponding author. Fax: +32 16 322 991.
E-mail addresses: jan.vanimpe@cit.kuleuven.be (J. Van Impe), geert.gins@cit.kuleuven.be (G. Gins).

http://dx.doi.org/10.1016/j.chemolab.2015.08.019
0169-7439/

SPM methods (such as those capable of dealing with inherent non-linearities of batch processes) can be applied in practice.

To properly assess the performance of various fault detection and identification methodologies, reliable and extended benchmarks are needed. For continuous processes, the Tennessee Eastman process published by Downs and Vogel [26] is widely used to benchmark various control and monitoring strategies [92,24]. Chiang et al. [17] published an extended reference set for fault detection and identification containing normal operation data and data from 22 different types of process upsets, available at http://web.mit.edu/braatzgroup/TE_process.zip. The relevance of a proper, extended benchmark is attested by the 157 citations of Downs and Vogel [26] indexed on Scopus in the period January 2014–May 2015 (17 months). Of these, 124 papers directly concern process monitoring.

When investigating the most important SPM techniques for batch processes as reviewed by Venkatasubramanian et al. [80]; Kourti [46]; MacGregor and Cinar [57]; Qin [63]; Aldrich and Auret [5], and Ge et al. [33], no benchmark comparable to the Tennessee Eastman process exists for batch processes, either in complexity (number of upsets) or frequency of use. Instead, most authors employ one or more small datasets.

For example, Nomikos and MacGregor [60,61] used a set of 51 normal and 2 faulty batches of a styrene-butadiene rubber (SBR) polymerization reaction generated with the model of Broadhead et al. [13] for their initial development of Multi-way Principal Component Analysis (MPCA) and Multiway Partial Least Squares (MPLS) for batch process monitoring. In Nomikos and MacGregor [62], they employed a set of 55 industrial two-stage polymerization batches provided by DuPont, of which 8 exhibit bad quality. The same DuPont dataset was used by Rännar et al. [65] to develop hierarchical PCA monitoring. Wold et al. [84] use data from an industrial fermentation to develop their alternative MPCA approach. Dahl et al. [23] employ data from 39 batch runs of an autoclave polymerization.

In their presentation of Batch Dynamic PCA (BDPCA) and Batch Dynamic PLS (BDPLS), Chen and Liu [15] used the SBR and DuPont datasets in addition to a set of 50 normal and 1 faulty batch of the CSTR problem originally presented by Luyben [56]. Choi et al. [21] also used the SBR dataset and a simulated batch MMA polymerization [1] of 100 normal and 3 faulty batches in the development of their autoregressive PCA (ARPCA) approach.

The SBR and DuPont datasets are also used in the review of Van Sprang et al. [78] and the comparison between global, evolving, and local PCA models for monitoring by Ramaker et al. [64]. These two papers also included three additional datasets: (i) an industrial multi-stage polymerization set of 47 normal and 3 abnormal batches (again provided by DuPont) [45], (ii) a collection of 67 normal and 3 faulty runs of an industrial batch polymerization of PVC [74], and (iii) a biochemical conversion set of 27 normal batches and 1 faulty batch [9]. Ramaker et al. [64] also employed 24 normal and 2 faulty batch runs of a fat hardening process originally presented by Smilde and Kiers [72] as a sixth dataset.

Lee et al. [50] generated 51 normal and 3 faulty batches using the Pensim simulated penicillin fermentation process of Birol et al. [10] to demonstrate SPM via Kernel PCA (KPCA). Jia et al. [43] used two datasets for Batch Dynamic KPCA (BDKPCA): a toy dataset (50 normal batches, 2 faulty) and Pensim (45 normal batches, 2 faulty).

The 2-dimensional DPCA (2D-DPCA) was developed by Lu et al. [55], Yao and Gao [89,90], and Yao et al. [88] using a toy problem, but the extensions towards Gaussian Mixture Model 2D-DPCA (GMM-2D-DPCA) [87], and 2-dimensional DKPCA and 2-dimensional Kernel Hebbian Algorithm (2D-KPCA and 2D-KHA) [98] are also tested on Pensim data of, respectively, 50 normal and 50 faulty batches, and 5 normal and 5 faulty batches.

Chen and Chen [14] used the Pensim (50 normal batches, 1 faulty) and SBR datasets to introduce Multi-Hidden Markov Tree-based MPCA (MHMT-MPCA) monitoring of batch processes. Zhao et al. [100] test Generalized Moving Window PCA (GMWPCA) via Pensim (20 normal batches) and an injection molding process (40 normal batches). Kulkarni et al. [48] combined PCA with Generalized Regression Neural Networks (PCA-GRNN), employing 48 normal and 4 faulty runs of the protein synthesis of Lim et al. [52] and 50 normal and 8 faulty batches of the penicillin production process of Lim et al. [53].

Recently, Multi-Scale PCA (MSPCA) for batch processes was proposed by Alawi et al. [2] and tested on 40 normal and 3 faulty Pensim batches.

Zhao and Shao [102], Zhang et al. [97], and Yu [95] all employed 100 normal and 3 faulty Pensim batches for their presentation of batch monitoring using, respectively Multiway Fischer Discriminant Analysis (MFDA), Kernel FDA (KFDA), and Multiway Kernel Localized FDA (MKLFDA). Yan et al. [86] proposed Semi-supervised Mixture Discriminant Monitoring (SMDM) as an improvement on MKLFDA using data form an injection molding process.

Lee et al. [49] and Yoo et al. [93] respectively generated 50 normal and 1 faulty, and 60 normal and 2 faulty Pensim batches to test SPM via Multi-way Independent Component Analysis (MICA). [3] conducted a more extensive test of MICA using Pensim (15 normal, 2 faulty) and DuPont datasets, and a third set of 40 normal runs and 1 faulty run of a simulated semi-batch production of polyol lubricant [96]. They later employ the same set of 15 normal and 2 faulty Pensim batches and the SBR dataset for Dynamic ICA (DICA) for batch monitoring [4]. Pensim was also used to generate 31 normal and 4 faulty batches for benchmarking Kernel ICA (KICA) by Tian et al. [75]. Ge and Song [31] developed a combined multilevel ICA-PCA methodology using the DuPont dataset. Zhao et al. [99] introduced combined Kernel ICA-PCA (KICA-PCA) employing data from Pensim (30 normal batches, 3 faulty) and from a three-tank system (18 normal batches, 2 faulty).

Zhao et al. [101] tested their dissimilarity measures for batch monitoring on a toy dataset and on 101 normal and 3 faulty Pensim batches. Hu and Yuan [41] generated 250 normal and 4 faulty Pensim batches for SPM by means of Tensor Locality Preserving Projections (TLPP) and also validated his procedure on 16 industrial batches. Alvarez et al. [6] used 187 normal and 444 faulty (8 types of faults at different magnitudes) Pensim for batch monitoring in the original measurement space—the largest Pensim dataset encountered by the authors.

An industrial dataset from a semiconductor etch process [83] consisting of 107 normal and 20 faulty batches is used in the works of Chen and Zhang [16] and Ge et al. [30,33] to respectively test Gaussian Mixture Models (GMM) and Support Vector Data Description (SVDD) for batch monitoring.

Fault identification for batch processes—if even discussed—mostly occurs after fault detection via analysis of contribution plots, despite their suffering from *fault smearing*, which possibly leading to incorrect diagnosis [82,77]. A few exceptions exist, such MKLFDA, where fault detection and identification occur simultaneously [95].

Classification models present an alternative approach to fault identification: given a set of known process upsets of various types, the model assigns the most probable cause to a detected new upset.

Cho and Kim [19,20] proposed an FDA-based fault classification using data from a simulated PVC polymerization. Hereto, they generated a set of 44 normal batches, and 3500 faulty batches of 5 types because their approach requires a number of faulty batches for classifier training greater than the dimensionality of the batches (in their case, the number of monitored sensors times the number of time points). Cho [18] tested a KFDA classifier for fault identification on two datasets: the same PVC polymerization and Pensim (60 faulty batches, 5 types). Li and Cui [51] also employ 60 faulty Pensim (5 types) in their work on Feature Vector Selection FDA using Nearest Feature Lines (FVS-FDA-NFL). No information is provided by Cho and Kim [19,20], Cho [18], or Li and Cui [51] on the type of the employed process upsets, their magnitude, or their onset time.

A total of 150 Pensim batches (50 of each of three types of process upsets) was used by Monroy et al. [59] to test fault identification via