UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA INDUSTRIAL - PEI

MESTRADO EM ENGENHARIA INDUSTRIAL

IZETE CELESTINA DOS SANTOS SILVA

AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS COM RECONCILIAÇÃO DE PADRÕES

SALVADOR
2018

UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA INDUSTRIAL

IZETE CELESTINA DOS SANTOS SILVA

AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS COM RECONCILIAÇÃO DE PADRÕES

SALVADOR/BA
2018

IZETE CELESTINA DOS SANTOS SILVA

AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS COM RECONCILIAÇÃO DE PADRÕES

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Industrial, da Universidade Federal da Bahia, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Industrial.

Orientadores: Dr. Cristiano Fontes Hora
Dr. Marcelo Embiruçu

Salvador/BA
2018

A553 Silva, Izete Celestina dos Santos

Agrupamento e classificação de séries temporais multivariadas com reconciliação de padrões / Izete Celestina dos Santos Silva. – Salvador, 2018.

112f.: il. color.

Orientador: Cristiano da Hora Fontes
Coorientador: Marcelo Santos Embiruçu

Dissertação (Mestrado em Engenharia Industrial) – Universidade Federal da Bahia, Escola Politécnica, 2018.

Referências: 55-58.

1.Agrupamento. 2.Fuzzy c-means. 3. Séries Temporais I. Cristiano Fontes. II. Universidade Federal da Bahia.

CDD.: 665.7

"AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS COM RECONCILIAÇÃO DE PADRÕES".

IZETE CELESTINA DOS SANTOS SILVA

Dissertação submetida ao corpo docente do programa de pós-graduação em Engenharia Industrial da Universidade Federal da Bahia como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Industrial.

Examinada por:

Prof. Dr. Cristiano Horas Fontes_____
Doutor em Engenharia Química, Brasil, pela Universidade Estadual de Campinas (Unicamp), 2001.

Prof. Dr. Eduardo Nascimento_____
Doutor em Física, pela Universidade de São Paulo (USP), São Paulo, 2004.

Prof. Dr. Flávio Morais de Assis Silva_____
Doutor em Informática, Alemanha, pela Universidade Technische Universitât Berlin, 1999.

Prof. Dr. Jorge Laureano Moya Rodriguez_____
Doutor em Projeto de Máquina, pela Universidade Central Marta Abreu de las Villas, Cuba, 1994.

Salvador, BA - BRASIL
Setembro/2018

“A mente que abre uma nova ideia
jamais volta ao seu tamanho
original”.
Oliver Wendell Holmes

Agradecimentos

- A Deus, pela oportunidade da minha existência na Terra.
- Aos meus pais, especialmente a meu pai, meu herói, meu amigo, meu tudo.
- Aos meus irmãos, em destaque a Lívia, a melhor irmã do mundo.
- A Pedro Aragão, uma eterna gratidão, por tudo que fez por mim, pelo aprendizado e pela nossa grande amizade.
- À minha eterna aluna Caterine, espero ter mudado um pouco a sua vida assim como você mudou a minha, obrigada pelas nossas discussões, que acrescentaram muito conhecimento.
- Ao meu grande amigo Reiner, nem tenho palavras para agradecer toda a sua paciência, todos os ensinamentos os constantes incentivos.
- Às minhas amigas Isabela, Jucileide, Luíza, Luciene e Patrícia pelo grande carinho que nutre nossa amizade.
- A Carla Galvão, Carla Mendes e Rafaela pela amizade nesses dois anos intenso de muita luta e vitórias.
- A Rodrigo, Harrison, Daniel, Ebert, Marcos, Carolina Amaro e Marcio pelo apoio, pelas discussões, amizade e pelos bons momentos de descontração.
- Meu orientador Cristiano, pela sua paciência, por todos o ensinamento compartilhados de uma forma singular, pela dedicação e confiança. Sou eternamente grata por tudo.
- Meu orientador Marcelo, pela oportunidade e confiança de ter me escolhido para realização desse trabalho.
- Raony pelo empenho da co-orientação neste trabalho, além da grande amizade.
- Aos meus eternos professores e amigos Fontoura, Lazaro e Janaína, pelos seus ensinamentos e experiências compartilhas, grata pela amizade.
- À Tatiane, Tamires, Robinson e Aragão, da secretária do PEI, pela competência em das inúmeras vezes que precisei, além da amizade.
- À Capes, pelo apoio financeiro.

Resumo da Dissertação apresentada ao PEI/UFBA como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS COM RECONCILIAÇÃO DE PADRÕES

Izete Celestina dos Santos Silva

Setembro/2018

Orientadores: Prof. Dr. Cristiano Hora Fontes

Prof. Dr. Marcelo Embiruçu

O armazenamento de uma grande quantidade de dados históricos de processos de produção estimulou o desenvolvimento de técnicas relacionadas à mineração de dados (*Data Mining*, DM) e à extração de conhecimento útil acerca do processo (*Knowledge Discovery in Data bases*, KDD). Embora existam muitos trabalhos relacionados à Detecção e Diagnóstico de Falhas (*Fault Detection and Diagnosis*, FDD), poucos deles são baseados em agrupamento e reconhecimento de padrões em séries temporais, especialmente em séries multivariadas. Além disso, na literatura revisada não há trabalhos relacionados ao reconhecimento de padrões em séries temporais multivariada que considerem o modelo de processo como restrição. À luz disso, este trabalho propõe um novo método para o reconhecimento de padrões em séries temporais uni e multivariada, baseado no algoritimo *Fuzzy C-Means* (FCM), que considera diretamente a dinâmica do processo no problema de agrupamento visando garantir, desta forma, a viabilidade dos padrões reconhecidos. O método proposto é aplicado em dois estudos de caso, ambos relacionados ao agrupamento e reconhecimento de padrões de operação anormal (falhas) e operação normal. O primeiro estudo de caso compreendeu um Reator Contínuo de Tanque Agitado (*Continuous Stirred Tank Reactor*, CSTR), que consiste em um processo de referência bem conhecido e utilizado para avaliar estratégias de controle e técnicas de FDD. A segunda aplicação envolveu um cenário industrial real que compreende uma turbina a gás, de escala comercial, localizada na unidade termoelétrica (UTE) Rômulo Almeida, parte integrante do parque da Companhia Brasileira de Petróleo. Os resultados obtidos evidenciam que o algoritmo FCM e uma métrica típica de similariedade entre séries temporais, baseada na Análise de Componentes Principais (PCA), não garantem o reconhecimento de padrões consistentes com a dinâmica do processo, mesmo com bons resultados de classificação e agrupamento. Por outro lado, os resultados obtidos a partir das abordagens de reconciliação propostas neste trabalho mostram a obtenção de padrões consistentes e reconciliados com a realidade dinâmica do processo, sem prejuízo da qualidade dos resultados de agrupamento e classificação.

Palavras-chave: Agrupamento, Fuzzy c-means, Reconciliação, Séries temporais.

Abstract of Dissertation presented to PEI/UFBA as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.).

CLUSTERING AND CLASSIFICATION OF MULTIVARIATE TEMPORARY SERIES WITH RECONCILIATION OF PATTERNS

Izete Celestina dos Santos Silva

Setembro/2018

Advisors: Prof. Dr. Cristiano Hora Fontes

Prof. Dr. Marcelo Embiruçu

The storage of a large amount of historical data in production processes has contributed to the development of techniques related to data mining (DM) and the extraction of useful knowledge about processes (Knowledge Discovery in Data bases, KDD). Although there are many studies related to Fault Detection and Diagnosis (FDD), few of them are based on grouping and pattern recognition in time series, especially in multivariate series. In addition, there are no work related to the recognition of patterns in time series that consider the process model as a constraint. This study proposes a new method for the recognition of patterns in uni and multivariate time series, based on the Fuzzy C-Means (FCM) algorithm, which directly considers the process dynamics in the clustering problem in order to guarantee the viability of the standards recognized. The proposed method is applied in two case studies, both related to clustering and recognition of patterns of abnormal operation (failures) and normal operation. The first case study is a Continuous Stirred Tank Reactor (CSTR), a well-known reference process used to evaluate control strategies and techniques for FDD. The second application involved a real industrial scenario comprising a commercial scale gas turbine located at the Rômulo Almeida thermoelectric plant (UTE), an integral part of the Companhia Brasileira de Petróleo park. The results show that the FCM algorithm and a typical metric of similarity between time series, based on the Principal Component Analysis (PCA), do not guarantee the recognition of patterns consistent with the process dynamics, even if good results are obtained classification and grouping. On the other hand, the results obtained from the reconciliation approaches proposed in this study show the obtaining of consistent and reconciled patterns with the dynamic reality of the process, without prejudice to the quality of the results of grouping and classification.

Keywords: Clustering, Fuzzy c-means, Reconciliation, Time series.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Símbolos

$\mathbf{\varepsilon}$	Coeficiente de fuzzificação
$\boldsymbol{\alpha}$	Parâmetro de sintonia
$\mathbf{A}$	Área da seção transversal do reator
$\mathbf{C_A}$	Concentração da espécie A no reator
$\mathbf{C_{AF}}$	Concentração da espécie A na corrente de alimentação
$\boldsymbol{\rho C_P}$	Capacidade de líquido refrigerante
$\boldsymbol{\rho_C}$	Capacidade calorífica da mistura
$\mathbf{C_{PC}}$	Capacidade calorífica do líquido refrigerante
$\boldsymbol{h}$	Nível do líquido no reator
$\mathbf{\Delta H}$	Calor de reação
$\mathbf{K_0}$	Fator pré-exponencial da equação de Arrhenius
$\mathbf{Q}$	Vazão volumétrica de saída do reator
$\mathbf{Q_C}$	Vazão volumétrica de líquido refrigerante
$\mathbf{E}$	Energia de ativação
$\mathbf{R}$	Constante universal dos gases
$\mathbf{T}$	Temperatura do reator
$\mathbf{T_C}$	Temperatura de líquido refrigerante da camisa
$\mathbf{T_{CF}}$	Temperatura de entrada do líquido refrigerante
$\mathbf{T_F}$	Temperatura da corrente de alimentação do reator
$\mathbf{U}$	Coeficiente de troca térmica
$\mathbf{A_C}$	Área da troca térmica
$\mathbf{C_{AF}}$	Concentração de entrada do reator
$\mathbf{Q_F}$	Vazão de alimentação no reator
$\mathbf{T_F}$	Temperatura do reagente
$\mathbf{Q_C}$	Vazão de alimentação da camisa
$\mathbf{T_{CF}}$	Temperatura do fluido refrigerante da camisa
$\mathbf{T}$	Temperatura do reagente dentro do reator
$\mathbf{Q}$	Vazão de saída do reator
$\mathbf{T_c}$	Temperatura de saída da camisa

$\mathbf{C_A}$	Concentração de saída da mistura
$\mathbf{T_E}$	Temperatura de Entrada
$\mathbf{T_S}$	Temperatura de Saída
$\mathbf{Q_A}$	Vazão de alimentação

Lista de Abreviaturas

FDD	Detecção e Diagnostico de Falha
KDD	*Knowledge Discovery in Data bases*
DM	*Data Mining*
FCM	*Fuzzy c-means*
CSTR	*Continuous Stirred Tank Reactor*
UTE	Unidade Termoelétrica
ARX	Modelo auto-regressivo com entrada externas
SPCA	Métrica de similaridade dos componentes principal
UST	Series Temporais Univariada
MST	Serie Temporal Multivariada
SPCA$_\lambda$	Métrica de similaridade dos componentes principais modificado
EP	Entradas Protótipos
EL	Entradas Livres

Sumário

CAPÍTULO 1

Introdução

Em décadas passadas, o desafio da indústria consistia em extrair informações baseadas em históricos de dados de processo, uma vez que havia pouca capacidade física de armazenamento e, ao mesmo tempo, disponibilidade de algoritmos ou técnicas capazes de obter conhecimentos úteis.

O simples armazenamento de dados em grande quantidade não implica na obtenção de informação útil. Neste sentido, técnicas específicas associadas à Mineração de Dados (*Data Mining*, DM) podem reconhecer aspectos relacionados aos padrões do comportamento que não seriam detectáveis por meio de técnicas estatísticas ou ferramentas analíticas tradicionais (SCHUCH; DILL, 2010). Em síntese, quanto maior e mais complexa a base de dados, mais difícil se torna a extração de conhecimentos úteis sobre o processo.

A análise de agrupamento de dados é uma das técnicas já consolidadas na área de mineração de dados (LIAO, 2005; BEZDEK; KELLER; KRISNAPURAM; PAL, 2005) e pode ser aplicada, entre outros fins, para o reconhecimento de padrões de operações normais ou anormais em processos de produção (SORSA; KOIVO, 1993). A análise de agrupamentos, por sua vez, pode ser utilizada para identificar ou reconhecer diferentes padrões de operação de uma planta industrial, a partir das informações contidas em dados históricos.

Dentre os métodos de agrupamento existentes, o algoritmo *Fuzzy C-Means* (FCM) representa uma alternativa robusta e consolidada de agrupamento não hierárquico, capaz de reconhecer padrões em séries temporais uni ou multivariadas associadas a variáveis de processo (LIAO, 2005; IZAKIAN; PEDRYCZ; JAMAL, 2015; TREBUŇA; HALČINOVÁ, 2013; D'URSO; MAHARAJ, 2012).

O agrupamento de séries temporais multivariadas é menos investigado pela complexibilidade dos desafios da escolha da métrica de similaridade e a avaliação da qualidade de classificação dos dados. Uma das métricas de similaridade amplamente utilizadas para a comparação entre séries temporais multivariadas é o fator de similaridade baseado na análise de componentes principais (*PCA similarity metrics*, SPCA) em suas

diferentes versões (SINGHAL; SEBORG, 2006; KHEDIRI; LIMAM; WEIHS, 2011; DENG; TIAN; CHEN, 2013).

A métrica SPCA e o algoritmo FCM clássico não são capazes de assegurar que os padrões reconhecidos sejam consistentes com o comportamento dinâmico do processo, o que, dependendo da qualidade das informações nos dados, pode levar à obtenção de padrões distantes da dinâmica dominante ou mesmo não realizáveis. Na literatura revisada, nenhum trabalho envolvendo a reconciliação de padrões em problemas de agrupamento de séries temporais multivariadas baseados em otimização foi identificado.

À luz dessas considerações, este trabalho revela, pela primeira vez, possíveis inconsistências no reconhecimento de padrões em séries temporais multivariadas usando a estratégia clássica do algoritmo FCM e define, também de forma inovadora, duas abordagens com diferentes resoluções envolvendo a reconciliação de padrões.

Inspirada na prática tradicional de reconciliação de dados (SHUANGHUA; MCLEAN; THIBAULT, 2007), a reconciliação de padrões significa modificar padrões obtidos a partir de dados históricos, tornando-os consistentes com a realidade do processo e, ao mesmo tempo, preservando a qualidade dos resultados de agrupamento/classificação.

Os métodos abordados neste trabalho são aplicados em dois estudos de caso. O primeiro é uma planta virtual (reator tanque contínuo agitado, CSTR) usada amplamente como referência para estudos de detecção e diagnóstico de falhas e estratégias de controle (SINGHAL; SEBORG, 2002). O segundo estudo de caso compreende a detecção de operação anormal (falha) na partida de uma turbina a gás em escala comercial em uma unidade termelétrica (Unidade Termelétrica Rômulo Almeida, Camaçari-Ba). Em ambos casos, os padrões identificados possibilitaram o reconhecimento da distinção entre os tipos de ocorrências de anormalidade.

Nesse panorama, este trabalho contempla tanto o uso de métodos tradicionais de agrupamento FCM com o fator de similaridade SPCA como o desenvolvimento de uma nova metodologia baseada no algoritmo FCM que acopla/uni o modelo do processo como restrição, capaz de reconhecer e classificar padrões em séries temporais multivariadas.

1.1 OBJETIVOS

1.1.1 Geral

Este trabalho objetiva apresentar possíveis inconsistências no reconhecimento de padrões em séries temporais multivariadas do algoritmo de agrupamento *fuzzy c-means*, além de propor dois métodos que consideram, ambos, a dinâmica do modelo do processo no problema de agrupamento para garantir a viabilidade dos padrões reconhecidos.

1.1.2 Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos foram determinados:

- Realizar uma revisão bibliográfica do estudo da arte sobre análise de agrupamento *fuzzy c-means* e fator de similaridade SPCA;
- Propor duas abordagens de reconciliação de padrões, sem prejuízo na qualidade da classificação;
- Implementar, usando o programa MatLab (*Simulink*), o algoritmo do modelo CSTR (Reator Tanque Agitado Continuo) para geração do banco de dados para testes;
- Testar e validar o algoritmo de agrupamento *fuzzy c-means* com reconciliação sequencial e simultânea.

1.2 JUSTIFICATIVA

A métrica de similaridade SPCA (Análise de Componente Principal) tradicional e o algoritmo de agrupamento FCM clássico não conseguem garantir que os padrões reconhecidos sejam consistentes com o comportamento dinâmico do processo, o que, dependendo da qualidade e quantidade das informações nos dados, pode levar à obtenção de padrões distantes da dinâmica dominante.

Na literatura revisada, nenhum trabalho envolvendo a reconciliação de padrões em problemas de agrupamento baseados em otimização foi verificado. Desta forma, é apresentado uma comparação entre o método FCM clássico e o algoritmo desenvolvido neste

trabalho. O algoritmo FCM mais que contempla a reconciliação dos padrões através do modelo dinâmico do processo.

1.3 ESTRUTURA DA DISSERTAÇÃO

Além deste capítulo, esta dissertação está dividida em cinco capítulos e um anexo.

O segundo capítulo traz fundamentos acerca das técnicas aplicadas e destaca os trabalhos relacionados ao tema encontrados na literatura.

O terceiro capítulo apresenta o delineamento metodológico desta pesquisa.

O quarto capítulo abrange os dois estudos de caso, nos quais a metodologia foi aplicada e os resultados obtidos, comparando-os com métodos tradicionais e os métodos desenvolvidos neste trabalho.

Por fim, o quinto capítulo apresenta a conclusão e algumas sugestões para futuros trabalhos.

CAPÍTULO 2

Análise de agrupamento

2.1 DEFINIÇÕES, OBJETIVO E CATEGORIAS DE MÉTODOS DE AGRUPAMENTO

A análise de *cluster* ou de agrupamento é aplicada em trabalhos de reconhecimento de padrões e tem como objetivo particionar um conjunto de dados em grupos, subconjuntos ou classes, permitindo a avaliação da dimensionalidade dos dados e identificação *outliers* (valores atípicos), isto é levantar hipóteses relacionadas à estrutura (associação) dos objetos (FÁVERO, 2009; JOHNSON; WICHERN, 2007; REIS, 2001). A análise de agrupamento é usada em diferentes campos como estatística, reconhecimento de padrões, aprendizado de máquina, mineração de dados, entre outros (JAIN; DUIN; MAO, 2000).

O amplo uso de algoritmos de agrupamento revela sua utilidade na análise exploratória de dados (JAIN; DUIN; MAO, 2000), além de ser uma ferramenta importante para analisar e revelar a estrutura (ou informação) muitas vezes oculta nos dados. A análise de agrupamento é uma técnica que utiliza a abordagem de aprendizagem não-supervisionada visando agrupar um conjunto de dados (objetos) de acordo com os princípios de homogeneidade (objetos pertencentes a um mesmo grupo) e heterogeneidade (objetos pertencentes a grupos distintos) (CULBERTSON; GURALNIK; STILLER, 2018; DÖRING; LESOT, 2006; HAIR; BLACK; BABIN; ANDERSON; TATHAN, 2009). A análise de agrupamento é uma etapa intrínseca no reconhecimento de padrões (BEZDEK; KELLER; KRISNAPURAM; PAL, 2005; HOPPNER; KLAWONN; KRUSE; RUNKLER, 1999; TREBUŇA; HALCINOVÁ, 2013).

Uma vez selecionado o conjunto de variáveis a serem analisadas, é necessário definir o método a ser utilizado no processo de agrupamento. Basicamente, há duas categorias de métodos de agrupamento de dados, que são, hierárquicos e não hierárquicos. Ambas as categorias analisam a distância entre indivíduos do mesmo grupo, indivíduos de grupos diferentes e a dispersão dos indivíduos dentro do mesmo grupo (REIS, 2001). Um método hierárquico é representado em uma estrutura de árvore (dendrograma) através da junção sistemática de grupos menores para formação de novos grupos. Um método não-hierárquico é baseado em um número predefinido de grupos e utiliza atributos dos dados para calcular a

proximidade de cada objeto ao centroide de cada agrupamento (LATTIN, 2011).

Os métodos não-hierárquicos são aplicados em diversos tipos de problemas e áreas do conhecimento, tendo em vista sua simplicidade, eficiência computacional e possibilidade de obtenção de bons resultados (LIAO, 2005). Dentre os métodos de agrupamento não hierárquico o *k-means* e o FCM são os algoritmos mais utilizados. A ideia principal destes algoritmos é a minimização da soma das distâncias de todos os objetos ao respectivo centro (ou padrão) de cada grupo de modo que a partição satisfaça dois requisitos básicos, quais sejam, "coesão" interna (ou semelhança interna) e isolamento (ou separação) dos grupos formados (MINGOTI, 2005; HAIR; BLACK; BABIN; ANDERSON; TATHAN, 2009).

Para objetos que podem ser representados na forma de vetor, a métrica de similaridade utilizada na versão clássicas dos algoritmos *k-means* e FCM é a distância euclidiana (Eq. 2.1).

$$\|x_i - x_j\| = \sqrt{(x_i - x_j)^T . (x_i - x_j)} \tag{2.1}$$

onde a distância entre os dois objetos $x_i \; e \; x_j \in \Re^p$.

O algoritmo *k-means* consiste em um processo iterativo de otimização alternada cuja finalidade é a minimização da soma das distâncias entre os objetos e os centroides de cada grupo (LIAO, 2005). O algoritmo se inicia com um número pré-definido de grupos e uma estimativa inicial (aleatória) para os respectivos centros. O problema de otimização descrito na Eq. 2.2, consiste em determinar a melhor distribuição dos objetos nos grupos e os melhores centros de cada grupo.

$$\min_{(V)} J(V) = \sum_{i=1}^{c} \sum_{k=1}^{n} \|x_k - v_i\|^2 \tag{2.2}$$

onde *c* é o número de grupos, *n* é o número de objetos. x_k é o *k-ésimo* do objeto, v_i é o centro ou padrão representativo do i-ésimo do grupo, e $\|\,.\,\|$ é a representação da métrica de similaridade.

O algoritmo *k-means* atribui cada objeto exclusivamente a um único grupo, resultando em um agrupamento rígido. Por outro lado, o algoritmo FCM parte do princípio de que cada objeto pode pertencer a mais de um grupo com diferentes valores de pertinência no intervalo [0,1], que quantifica o grau de associação ou aderência de cada objeto a cada um dos grupos (BERGET; MEVIK; NEAS, 2007; DUNN, 1973; MANGIAMELLI et al., 1996 *apud* MIGINOT, 2005; BEZECK, 1981; MEMON; LEE, 2018; KESEMEN; TEZEL; OZKUL, 2016).

Ao contrário do algoritmo *k-means*, o algoritmo FCM constitui-se uma ferramenta eficiente para descrever a inerente incerteza na definição das fronteiras entre os grupos (APARAJEETA; NANDA; DAS, 2016), sendo capaz de representar a estrutura intrínseca aos dados através de um modelo de agrupamento mais realista (DÖRING; LESOT, 2006). A família do método *c-means* é composta por três conjuntos de algoritmo que produzem partições diferentes de dados: rígido, probabilístico e possibilístico (BEZDEK, 1981). A abordagem probabilística é a mais utilizada na literatura (BEZDEK, 1981; DÖRING; LESOT, 2006; DUNN, 1973), sendo adotada também neste trabalho.

2.2 ALGORITMO FUZZY C-MEANS PROBABILÍSTICO

O algoritmo FCM foi desenvolvido por Dunn (1973) e aperfeiçoado por Bezdek (1981). Este é um método de agrupamento não hierárquico, baseado em otimização, que possui a capacidade de tratar as incertezas e sobreposições inerentes ao problema de agrupamento (CHIU, 1994; ABONYI; FEIL, 2007). Na área de reconhecimento de padrões, o FCM mostra-se muito útil em problemas envolvendo amostras com pouca informação de classes e dificuldade de definição de fronteiras entre os grupos (BEZDEK; KELLER; KRISNAPURAM; PAL, 2005; MENDEL, 2001; ZADEH, 1965, 1999).

O termo *fuzzy* decorre simplesmente da inerente característica do FCM em atribuir níveis de pertinências intermediários (no intervalo [0;1]) para cada objeto em relação à cada um dos grupos. A partição *fuzzy* é obtida como resultado do próprio problema de agrupamento, sendo representada através de uma matriz de partição que indica o nível de aderência de cada objeto a cada um dos grupos reconhecidos (ZADEH, 1965; BEZDEK, 1981; BERGET; MEVIK; NEAS, 2007; HOPPNER; KLAWONN; KRUSE; RUNKLER,

1999).

Sendo um método não hierárquico, o FCM requer a pré-especificação do número de grupos (BEZDEK, 1981) e compreende o problema de otimização descrito na Eq. 2.3, cujas variáveis de decisão são os centros V (v_i, i = 1, ..., c, centros ou padrões) de cada um dos grupos c ($c \geq 2$) e o grau de pertinência de cada objeto a cada grupo (LIAO, 2005; BEZDEK; KELLER; KRISNAPURAM; PAL, 2005).

$$\min_{(U,V)} J_\varepsilon(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left\{u_{ik}^{\varepsilon}\,\|\chi_y - v_i\|^2\right\} \tag{2.3}$$

onde c é o número de grupos, n é o número de objetos, u_{ik} é o grau de pertinência do *k-ésimo* objeto ao *i-ésimo* grupo e $\boldsymbol{U}$ é a matriz de partição (c x n, matriz). O parâmetro ε (ε > 1) é o coeficiente fuzzificador (recomendado na literatura ε = 2) e está relacionado ao nível de incerteza do problema de partição. $\boldsymbol{V}$ é o conjunto de vetores dos centros ou padrões {v_1, v_2, ..., v_c}.

Duas restrições adicionais devem ser consideradas (Eq. 2.4 e Eq. 2.5):

$$u_{ij} \in [0,1] \quad e \quad \sum_{j=1}^{n} u_{ij} > 0 \;\forall i \in \{1,\dots,c\} \tag{2.4}$$

$$\sum_{i=1}^{c} u_{ij} = 1 \;\forall j \in \{1,\dots,n\} \tag{2.5}$$

A restrição da Eq. 2.4 garante que nenhum grupo ficará vazio e a Eq. 2.5 estabelece que a soma das pertinências de um objeto, a todos os grupos, deve ser igual a unidade (abordagem probabilística) (BEZDEK, 1981).

Bezdek (1981) sugere a escolha do expoente de fuzzificação (ε) no intervalo [1,1; 5]. Na medida em que $\varepsilon \to 1$, o resultado do agrupamento tende ao caso rígido (*crisp*) com uma das pertinências de cada objeto aproximando à unidade e as demais chegando a zero. Quando $\varepsilon \to \infty$ o nível de incerteza do agrupamento se eleva e os objetos possuir todas as pertinências

próximas de 0,5 (BERGET; MEVIK; NEAS, 2007).

O algoritmo de agrupamento FCM é iniciado no momento em que são atribuídos estimativas iniciais para os centros ou protótipos de cada um dos grupos (Eq. 2.3). O FCM é um algoritmo não-supervisionado no qual os objetos não são rotulados previamente (BEZDEK, 1981) e as estimativas iniciais dos centros podem ser computadas através de médias envolvendo subconjunto de objetos aleatoriamente definidos.

A aplicação das condições de primeira ordem ao problema descrito pelas equações 2.3-2.5 produz a seguinte solução analítica para o método FCM (Eq. 2.6 e Eq. 2.7) (CHUANG et al., 2006):

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n}(u_{ik})^{\varepsilon} \cdot \chi_y}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} \qquad i = 1, \ldots, c \qquad (2.6)$$

e

$$u_{ik} = \frac{\left(\frac{1}{\|\chi_y - \boldsymbol{v}_i\|^2}\right)^{\frac{1}{\varepsilon-1}}}{\left(\frac{1}{\|\chi_y - \boldsymbol{v}_j\|^2}\right)^{\frac{1}{\varepsilon-1}}} \qquad k = 1, \ldots, n \qquad i = 1, \ldots, c \qquad (2.7)$$

O algoritmo FCM clássico compreende um procedimento iterativo envolvendo as equações 2.6 e 2.7. A solução analítica para determinação dos centros de cada grupo (Eq. 2.6) somente é válida quando a métrica de similaridade adotada é a distância Euclidiana. Portanto, em situações nas quais esta métrica não é utilizada ou mesmo no caso de objetos que não possam ser representados como uma grandeza vetorial (no caso de séries multivariadas), a equação 2.6 não se aplica e o método FCM deve ser resolvidos através da resolução numérica do problema de otimização é definido pelas equações 2.3 a 2.5. Este procedimento viabiliza perfeitamente a aplicação do FCM em casos mais complexos como, por exemplo, agrupamento de séries multivariadas.

Em resumo o algoritmo FCM clássico (Eq.s 2.6 e 2.7) compreende as seguintes etapas descritas no Quadro 1.

Quadro 1 – Algoritmo *fuzzy c-means*

Etapas	Descrição
1	Determinação de estimativa inicial para os centros de cada grupo;
2	Atualização da matriz de partição através da Eq. 2.7;
3	Obtenção de novos centros através da Eq. 2.6;
4	O processo iterativo é encerrado quando a mudança dos padrões é menor que o valor estabelecido (predeterminado) no critério de tolerância.

2.3 SERIES TEMPORAIS E A MÉTRICA SPCA (PCA)

As séries temporais univariadas (STU) e multivariadas (STM) são objetos bastante empregados em problemas envolvendo análise de agrupamento e reconhecimento de padrões (D'URSO; MAHARAJ, 2012; LI; WEN, 2014).

As séries temporais compreendem uma conjunto de observações ao longo do tempo associadas a uma variável de processo específica. Seja *p* o número de variáveis, *m* o número de observações/medições associados à cada variável e *t* um instante de tempo qualquer. Uma STU é aquela na qual $p = 1$ e quando $p \geq 2$ tem-se uma STM. Uma STM pode ser representada pela seguinte matriz de dimensão $m \times p$ (Eq 2.8; FONTES; BUDMAN, 2017).

$$Z_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ip}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ip}(m) \end{bmatrix} \tag{2.8}$$

onde Z_i é o objeto, z_{ij} (t) é a medida da variável j (j = 1, ..., p) no instante instantâneo t (t = 1, ..., m) no objeto Z_i (i = 1, n objetos). A coluna j contém a série temporal relacionada à variável j no objeto Z_i.

As séries temporais associadas à cada variável em uma STM devem ser consideradas de forma integrada e o comportamento ou característica do objeto multivariado não deve ser extraído analisando-se isoladamente cada uma das séries (O'REILLY; MOESSNER; NATI, 2017; TAK-CHUNG FU; 2011; YANG; SHAHABI; 2004; LIAO, 2005).

O índice de similariedade SPCA é baseado na PCA (YANG; SHAHABI, 2004), sendo uma métrica consolidada para a quantificação de distância entre STM. A PCA é uma técnica da estatística multivariada aplicada em diversas áreas: química analítica (SUCHACZ, 2010),

geologia (NOWICKI; ŻYLIŃSKA; KIN, 2013), agricultura (KOLASA-WIECECK, 2012), psicologia e sociologia (BRZYSKI; TOBIASZ-ADAMCZYK; KNUROWSKI, 2012; RASKIN; TERRY., 1988), controle de qualidade de alimentos (CZERNYSZEWICZ, 2008; RYMUZA et al., 2013), imagem e processamento de sinais (HLADNIK, 2013). O PCA oferece uma alternativa de redução de dimensionalidade preservando a variabilidade dos dados da amostra original.

O PCA consiste em determinar um conjunto de vetores ortogonais (*loading vectors*) através da decomposição em autovalores/autovetores da matriz de covariância. Desta forma, o PCA inicia considerando uma matriz de dados $X \in \Re^{m \times p}$ de p variáveis e m indivíduos. A matriz de convariância de X representa uma maneira útil de obter todos os valores possíveis entre as diferentes variáveis medidas, sendo definida pela Eq 2.9 (SINGHAL; SEBORG, 2002):

$$S = \frac{1}{m\ -1}\ .X^T\ X\ =\ V\ .\Lambda\ .V^T \tag{2.9}$$

Onde:

$V\ \in\ \Re^{pxp}$ colunas são ortonormais

$\Lambda\ \in\ \Re^{pxp}$ matriz diagonal com os autovalores reais não negativos em ordem decresente ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$)

cada autovalor λ_i (i=1, ..., p) representa a variância da amostra original projetada na direção do *n-ésimo* componente (*n-ésima* coluna da matriz V).

Um conjunto de componentes principais (y) é capaz de representar ou explicar um percentual da variância total da amostra original dos dados. A redução de dimensionalidade consiste na seleção dos autovetores ou componentes principais capazes de representar um percentual acima de 95% da variabilidade dos dados originais (YANG; SHAHABI, 2004; HAIR; BLACK; BABIN; ANDERSON; TATHAN, 2009; FONTES; BUDMAN, 2017).

Dada uma quantidade y de componentes principais que representa pelo menos 95% da variabilidade da amostra original ($y<p$), é possível projetar as variáveis originais em um novo espaço de menor dimensão.

O índice de Similaridade PCA quantifica a similaridade entre duas séries temporais multivariadas através da comparação entre as direções dos respectivos componentes

principais (SINGHAL; SEBORG, 2002; YANG; SHAHABI, 2004). O índice é limitado ao intervalo [0; 1], sendo que os valores próximos de 1 indicam alta similaridade e valores próximo a 0 alta dissimilaridade (Eq. 2.10).

$$SPCA(Z_A, Z_B) = \frac{1}{k_0} . \sum_{i=1}^{k_0} \sum_{j=1}^{k_0} . (cos\theta_{ij})^2 \tag{2.10}$$

onde: Z_A e Z_B são as matrizes que representam os objetos. θ_{ij} é ângulo entre o *i-ésima* componente principal de Z_A e o *j-ésimo* componente principal de Z_B. k_0 é definido como maior valor entre o k_A e k_B (número de componentes principais de Z_A e Z_B, respectivamente).

No índice SPCA tradicional [Eq. (2.10)] os componentes principais possuem o mesmo peso, entretanto, essas mesmas componentes representam diferentes percentuais de variabilidade quantificada de acordo com o respectivo autovalor. Desta forma, foi proposto o índice SPCA modificado ($SPCA_\lambda$) no qual cada componente principal é ponderado pelo seu autovalor correspondente (LI; WEN, 2014; DENG; TIAN; CHEN, 2013; SINGHAL; SEBORG, 2006):

$$SPCA_c(Z_A . Z_B) = \frac{1}{\sum_{i=1}^{k_0} (\lambda_i^A . \lambda_i^B)} . \sum_{i=1}^{k_0} \sum_{j=1}^{k_0} \left(\lambda_i^A . \lambda_j^B\right) . (cos\theta_{ij})^2 \tag{2.11}$$

onde: λ^A e λ^B compreendem os autovalores de $Z_A^T . Z_A$ e $Z_B^T . Z_B$, respectivamente. θ_{ij} é o ângulo entre o *i-ésimo* componente principal de Z_A e o *j-ésimo* componente principal de Z_B.

O uso do PCA para fins de cálculo do SPCA entre duas séries temporais multivariadas requer que cada uma das colunas (variáveis) da amostra original (matriz *X*) seja centralizada na média o que é feito através da subtração de cada valor da respectiva média aritmética entre todos os valores da mesma coluna. A aplicação do SPCA na análise de similaridade de séries multivariadas requer que os objetos tenham o mesmo número de variáveis, mas não necessariamente o mesmo número de medições (ou comprimento da janela de tempo) (SINGHAL; SEBORG, 2002).

2.4 ALGORITMO FUZZY C-MEANS MODIFICADO PARA SÉRIES TEMPORAIS MULTIVARIADAS

O algoritmo FCM clássico compreende o uso da distância Euclidiana como métrica de similaridade/dissimilaridade, sendo válido somente para agrupamento de séries temporais univariadas (BERGET; MEVIK; NEAS, 2007; ZHANG, 2017). O processo é iterativo é a resolução analítica compreende as equações 2.6 e 2.7. A distância Euclidiana não se aplica a séries temporais multivariadas, desta forma, não é possível a determinação analítica dos centros de cada grupo, sendo necessário um método numérico para resolução do problema de otimização. Neste trabalho a métrica de similaridade utilizada no algoritmo FCM é a SPCA e compreende o seguinte problema de otimização, visto na Eq. 2.12 (FONTES; PEREIRA, 2016):

$$\underbrace{min}_{U,V}\ \Omega_{\varepsilon}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon}\cdot\left(SPCA_c(X_k,V_i)\right)^2\right) \tag{2.12}$$

Sujeito à Eq. 2.13:

$$\begin{cases} u_{ik} \in [0,1] \quad and \quad \sum_{k=1}^{n} u_{ik} > 0 \ \ \forall i \\ \sum_{i=1}^{c} u_{ik} = 1 \ \ \forall k \end{cases} \tag{2.13}$$

Onde:

$$SPCA_c(X_k,V_i) = 1 - SPCA_{\lambda}(X_k,V_i) \tag{2.14}$$

O $SPCA_c$ é o complemento do $SPCA_{\lambda}$ (Eq. 2.14), uma vez que valores de $SPCA_{\lambda}$ próximos da unidade implicam em elevada similaridade. V é o conjunto dos centros ou padrões, $V_i(i = 1, \dots, c)$ e X_k $(k = 1, \dots, n)$ são os objetos (séries temporais multivariadas) $\left(X_k\ e\ V_i\ \in\ \Re^{m\times p}\right)$. $SPCA_c(X_k,V_i)$ é a distância entre os objetos e o centro de cada grupo com base no SPCA modificado (Eqs. 2.12-2.14).

A aplicação das condições de primeira ordem para a Eq. 2.12 produz a Eq. 2.15 para a determinação dos graus de pertinência (matriz de pertinência) de cada objeto a partir dos

centros de cada grupo.

$$u_{ij} = \frac{\left(\frac{1}{(SPCA_c(X_k,V_i))^2}\right)^{\frac{1}{\varepsilon-1}}}{\sum_{k=1}^{c}\left(\frac{1}{(SPCA_c(X_k,V_i))^2}\right)^{\frac{1}{\varepsilon-1}}} \tag{2.15}$$

O algoritmo FCM e o métrica $SPCA_\lambda$ apresentam uma tendência a convergir para mínimos locais nos quais os protótipos associados à cada grupo são muito próximos entre si. Como consequência disso, o resultado final do agrupamento terá todos os objetos pertencentes a um único grupo.

A fim de evitar este resultado indesejável, foi proposta a adição de uma nova parcela à função objetivo (Eq. 2.12), que é inversamente proporcional à soma das distâncias entre os protótipos de cada grupo. A inclusão deste novo critério na função objetivo procura maximizar a divisão entre grupos (DAO; DUONG; VRAIN, 2017), evitando a obtenção de centros muito similares.

$$\underbrace{min}_{U,V}\ \Omega_\varepsilon(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon}\cdot\left(SPCA_c(X_k,V_i)\right)^2\right) + \beta\cdot\sum_{i=1}^{c}\sum_{\substack{j=1\\ j>i}}^{c}\frac{1}{\left(SPCA_c\left(V_j,V_i\right)\right)^2} \tag{2.16}$$

onde: β é um fator ajustável que determina o peso da parcela adicionada.

Foi verificado que valores pequenos (da ordem de 10^{-5}) são suficientes para evitar a convergência do agrupamento a protótipos similares, sem prejudicar a qualidade final do agrupamento. A Eq. 2.16 define um problema de agrupamento com bi-critério que evita a possibilidade de obtenção de padrões muito próximos (ou mesmo similares), o que implicaria na sobreposição de grupos.

CAPÍTULO 3

Reconciliação de padrões

A reconciliação de dados é uma técnica bastante difundida e tem sido aplicada em diferentes tipos de problemas (SHUANGHUA; MCLEAN; THIBAULT, 2007). A reconciliação de dados consiste em minimizar o erro ou desvio entre o valor medido de uma variável do processo e o valor predito para esta variável através de um modelo do próprio processo. Em outras palavras, reconciliar dados significa aproximar medições que estão sujeitas a ruídos da fenomenologia de um processo, tornando-as consistentes com a realidade física do mesmo. Podemos citar alguns exemplos de aplicações como na qualidade dos dados em um usina termoelétrica de carvão, pela baixa precisão dos instrumentos de medição (JIANG; LUI; LI, 2014), aprimoramento e monitoramento de padrões de uma turbina a gás a fim de diagnosticar a existência de operação anormal do equipamento (SYES; DOOLEY; MADRON; KNOPF, 2016), reconciliação de dados de um condensador de arrefecimento com o objetivo de identificar o desempenho real do equipamento sob diferentes situações (LI et al., 2018).

Até setembro de 2018, data de defesa desta dissertação, verificou-se na literatura revisada que não existem trabalhos sobre a reconciliação de padrões envolvendo agrupamento e classificação de séries temporais. Este capítulo apresenta o problema de reconciliação de padrões em séries temporais, define os tipos de problemas e propõe as respectivas abordagens de resolução.

3.1 SÉRIES TEMPORAIS MULTIVARIADAS (STM) APENAS COM VARIÁVEIS DE SAÍDA

O primeiro problema compreende um cenário no qual todas as séries temporais presentes em cada objeto são variáveis de saída do processo, ou seja, não são manipuláveis e representam o efeito de outras variáveis (entradas). Neste caso, a reconciliação dos padrões

obtidos consiste em determinar a trajetória temporal de cada variável de saída considerada mais próxima do padrão reconhecido pelo algoritmo de agrupamento mas que, ao mesmo tempo, seja consistente com a dinâmica do modelo processo. A resolução deste problema compreende a determinação de uma seqüência de valores das entradas consistentes com a realidade (satisfaz as restrições físicas do processo) e, como consequência, a trajetória respectiva da variável de saída predita pelo modelo de processo (padrão reconciliado).

Uma vez que os centros ou padrões são reconhecidos através do algoritmo de agrupamento FCM (V_i, *i = 1, ..., c,* [Eq. 2.16-3.1]), a reconciliação de cada padrão reconhecido compreende o seguinte problema de otimização, a ser resolvido de forma sequencial, posteriormente ao problema de agrupamento (Eq. 3.2). Este problema de otimização visa aproximar o perfil dinâmico de cada variável de saída (série temporal) às caracteristicas do processo, com base no seu modelo dinâmico e das restrições físicas.

$$V_i = \begin{bmatrix} v_{i1}(1) & \cdots & v_{ip}(1) \\ \vdots & \ddots & \vdots \\ v_{i1}(m) & \cdots & v_{ip}(m) \end{bmatrix} = \{V_{i1},\ V_{i2}, \dots ,\ V_{ip}\} \tag{3.1}$$

onde cada V_{ij} é uma série temporal relacionada à variável j $(j = 1, \dots, p)$ no padrão i $(i = 1, \dots, c)$ $\left(V_{ij} \in \Re^m\right)$, tem-se o seguinte problema de otimização a ser resolvido após a aplicação do método do algoritmo FCM (Eqs 3.2-3.7).

$$\min_{W,V^r} \Pi_\varepsilon\,(V^r, W) = \sum_{i=1}^{c}\sum_{j=1}^{p}\left\|V_{ij}^r - \hat{\gamma}_{cij}\right\|^2 \tag{3.2}$$

Sujeito a

$$\hat{y}_{c_{ij}} = \left[\hat{y}_{c_{ij}}(1), \dots, \hat{y}_{c_{ij}}(m)\right]^T \tag{3.3}$$

$$\hat{y}_{c_{ij}}(k_t) = f_{dj}\left(w_{i1}(k_t - dt_1), \dots, w_{i1}(k_t - nw_1), w_{i2}(k_t - dt_2), \dots, w_{i2}(k_t - nw_2), \dots, w_{ig}\left(k_t - dt_g\right), \dots, w_{ig}\left(k_t - nw_g\right)\right)$$

$$k_t = 1, \dots, m;\ \ i = 1, \dots, c\ \ e\ \ j = 1, \dots, p \tag{3.4}$$

$$L_l \leq w_{il}(k_t) \leq S_l, \quad l = 1, \dots, g \;\; and \;\; i = 1, \dots, c \tag{3.5}$$

$$\Delta L_l \leq \Delta w_{il}(k_t) \leq \Delta S_l, \;\; onde \;\; \Delta w_{il}(k_t) = w_{il}(k_t) - w_{il}(k_t - 1) \tag{3.6}$$

$$l = 1, \dots, g \quad e \quad i = 1, \dots, c \tag{3.7}$$

onde: f_{dj} é o modelo do processo na forma discreta que relaciona a saída j $(j = 1, \dots, p)$ com as entradas do processo, $dt_l \; e \; nw_l$ $(l = 1, \dots, g)$ são o tempo morto e número de valores passados, respectivamente, de cada entrada no modelo discreto. k_t é o instante de tempo, $L_l \; e \; S_l$ são os limites inferior e superior de cada entrada e $\Delta L_l \;\; e \;\; \Delta S_l$ são os limites inferior e superior para as variações de cada entrada em cada instante de tempo (representam pois restrições físicas associadas às variáveis de entrada). $w_{il}(k_t)$ é o valor da variável de entrada *l* no instante de tempo k_t, associada ao grupo i $(i = 1, \dots, c)$. W é o conjunto de perfis de entrada associados à cada grupo $\{w_{il} \in \Re^m, i = 1, \dots, c \;\; e \;\; l = 1, \dots, g\}$ e V^r é o conjunto (STM) de padrões reconciliados $\left(V_{ij}^r \in \Re^m, i = 1, \dots, c, \; j = 1, \dots, p\right)$. $\hat{y}_{cij}$ $\left(\hat{y}_{cij} \in \Re^m\right)$ é a resposta dinâmica da saída *j* mais próxima de seu padrão reconciliado no grupo *i*.

O padrão reconciliado é aquele resultante de um conjunto de trajetórias temporais realizáveis das variáveis de entrada e que, concomitantemente, aproxima se do padrão reconhecido pela técnica de agrupamento. A métrica de distância usada na Eq. 3.2 é a euclidiana, uma vez que essa métrica é capaz de considerar o formato ou dinâmica da série temporal, o que não ocorre com o SPCA.

3.2 SÉRIES TEMPORAIS MULTIVARIADAS (STM) COM VARIÁVEIS DE ENTRADA E SAÍDA

O segundo tipo de problema compreende séries temporais multivariadas envolvendo variáveis de saída e variáveis de entrada do processo. Nesse caso, a reconciliação consiste em relacionar as séries temporais referentes às variáveis de saída, em cada padrão reconhecido, com as séries temporais referentes às entradas no mesmo padrão, através do modelo de processo.

Considera-se um processo com um total de g_1 variáveis de entradas $(g_1 < g)$ contempladas em cada objeto (STM) e g_2 entradas restantes $(g_1 + g_2 = g)$. As primeiras são denominadas de Entradas Protótipo (*EP*) e as segundas são denominadas de Entradas Livres (*EL*). Desta forma, o conjunto de variáveis de entrada se divide em dois subconjuntos, quais sejam, w_j^{EP}, $(j = 1, \dots, g_1)$ e w_j^{EL}, $(j = 1, \dots, g_2)$. Cada padrão reconhecido V_i $(i = 1, \dots, c)$ e cada objeto é composto por séries temporais referentes a p saídas e g_1 entradas protótipos e compreende uma matriz com $p + g_1$ colunas. $V_{ij}^{y} \in \Re^{m}$ $(i = 1, \dots, c$ e $j = 1, \dots, p)$ refere-se à serie temporal da *j-ésima* saída no padrão associado ao cluster i e w_{ij}^{EP} $(i = 1, \dots, c$ e $j = 1, \dots, g_1)$ refere-se à série temporal da *j-ésima EP* no padrão associado ao *cluster i* (Eq. 3.8).

$$V_i = \left\{ \underbrace{V_{i1}^{y}, \dots, V_{ip}^{y}}_{p\ sa\acute{i}da} ; \underbrace{w_{i1}^{PI}, \dots, w_{ig_1}^{PI}}_{\substack{g_1\ padr\tilde{o}es\ de \\ entrada}} \right\} \tag{3.8}$$

Neste caso, a estratégia desta reconciliação é realizada através de uma abordagem simultânea. O problema de otimização consiste em uma função objetivo cuja primeira parcela está relacionada ao agrupamento FCM (Eq. 3.9) propriamente dito e a segunda parcela estabelece uma ligação entre as séries temporais de cada padrão com base no modelo do processo. Cada padrão possui séries temporais de variáveis de saída e variáveis de entrada e há necessariamente um relação entre elas definidas pelo modelo. O problema de otimização é estruturado conforme a Eq. 3.9.

$$\underbrace{min}_{U,V,W}\ \rho_{\varepsilon}(U,V,W) =$$

$$\alpha \cdot \left(\underbrace{\sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon} \cdot \left(SPCA_c(X_k, V_i)\right)^2\right) + \beta \cdot \sum_{i=1}^{c}\sum_{\substack{j=1 \\ j>i}}^{c} \frac{1}{\left(SPCA_c(V_j, V_i)\right)^2}}_{Agrupamento} \right) \tag{3.9}$$

$$+\ \underbrace{(1-\alpha) \cdot \sum_{i=1}^{c}\sum_{j=1}^{p}\left\| V_{i1}^{y} - \hat{y}_{c_{ij}} \right\|^2}_{padr\tilde{o}es\ reconciliados}$$

Sujeito às seguintes restrições,

$$\begin{cases} u_{ik} \in [0,1] \;\; and \;\; \sum_{k=1}^{n} u_{ik} > 0 \;\; \forall i \\ \sum_{i=1}^{c} u_{ik} = 1 \;\; \forall k \end{cases} \tag{3.10}$$

$$\hat{y}_{c_{ij}} = \left[\hat{y}_{c_{ij}}(1), \dots, \hat{y}_{c_{ij}}(m)\right]^T \tag{3.11}$$

$$\begin{aligned} \hat{y}_{c_{ij}}(k_t) = \; & f_{dj}\Big(w_{i1}^{EP}(k_t - dt_1^{EP}), \dots, w_{i1}^{PI}(k_t - nw_1^{EP}), , \dots, w_{ig_1}\big(k_t \\ & - dt_{g_1}^{EP}\big), \dots, w_{ig_1}\big(k_t - nw_{g_1}^{EP}\big), w_{i1}^{EL}(k_t - dt_1^{EL}), \dots, w_{i1}^{EL}(k_t \\ & - nw_1^{EL}), , \dots, w_{ig_2}\big(k_t - dt_{g_2}^{EL}\big), \dots, w_{ig_2}\big(k_t - nw_{g_2}^{EL}\big)\Big); k_t \\ & = 1, \dots, m; \;\; i = 1, \dots, c \;\; e \;\; j = 1, \dots, p \end{aligned} \tag{3.12}$$

$$L_l \leq w_{il}(k_t) \leq S_l, \quad l = 1, \dots, g \;\; e \;\; i = 1, \dots, c \tag{3.13}$$

$$\Delta L_l \leq \Delta w_{il}(k_t) \leq \Delta S_l, \;\; onde, \;\; \Delta w_{il}(k_t) = w_{il}(k_t) - w_{il}(k_t - 1)\,, \tag{3.14}$$

$$l = 1, \dots, g \quad e \quad i = 1, \dots, c \tag{3.15}$$

As restrições físicas (Eqs. 3.14-3.15) aplicam-se a todas as entradas, independentemente da sua categoria (EL ou EP). Cada modelo deve considerar o efeito de todas as entradas (Eq. 3.12). dt_j^{PI} *e* nw_j^{PI} $\left(dt_j^{NPI} \text{ e } nw_j^{NPI}\right)$ são o tempo morto e o número de valores da *j-ésima EP* do modelo do processo.

Assim como no problema anterior (abordagem de reconciliação sequencial), *W* é o conjunto de perfis de entrada associada à cada grupo $\{w_{il} \in \Re^m, i = 1, \dots, c \;\; e \;\; l = 1, \dots, g\}$. O comportamento dinâmico de EL pode ser definido ou não na otimização (Eqs. 3.9-3.15), dependendo das informações relativas ao problema real. Se forem pré-fixadas, as EL não serão variáveis de decisão. Na segunda parcela da Eq. (3.9) é utilizada a distância euclidiana, pela sua capacidade de considerar a dinâmica da série temporal. O $\alpha(\alpha \in [0,1])$ é um parâmetro de sintonia (parâmetro de peso das parcelas). v_{iy} é a série temporal para a saída *y* (y = 1, ..., n_y), associada ao padrão do grupo *i* (i = 1, ..., c).

O parâmetro da sintonia (α) visto na Eq. 3.9 determina o peso que o modelo terá no problema de otimização. Desse modo, atribuir valores elevados de α implica em considerar pouca influência do modelo no problema de agrupamento.

CAPÍTULO 4

Resultados e discussão

Neste capítulo é feito um comparativo dos padrões resultantes do algoritmo FCM e do algoritmo de reconciliação (sequencial e simultâneo) através do banco de dados históricos gerados por dois estudos de caso. O primeiro é uma unidade virtual de um reator tanque agitado contínuo (CSTR) e o segundo um cenário real de uma turbina a gás com o objetivo de reconhecer padrões em séries temporais multivariadas em problema envolvendo detecção de falhas.

4.1 ESTUDO DE CASO 1: DETECÇÃO DE FALHAS EM UM REATOR TANQUE AGITADO CONTÍNUO

4.1.1 Descrição do Processo

O primeiro estudo de caso compreende um Reator de Tanque Agitado Contínuo não-isotérmico (CSTR) com dinâmica da camisa de refrigeração e nível de mistura (conteúdo reacional) variável. O CSTR é um processo de referência comum em abordagens de FDD, principalmente reações em fase líquida, operando em regime contínuo ou em batelada (SINGHAL; SEBORG, 2002; VAIDYANATHAN; VENKATASUBRAMANIAN, 1992). Duas malhas de controle são consideradas, quais sejam, na malha de controle em cascata da temperatura do reator cuja variável manipulada é a vazão de líquido refrigerante na camisa e uma malha de controle de nível cuja variável manipulada é a vazão de saída do reator. A estrutura e os parâmetros da malha de controle em cascata utilizada, assim como as condições operacionais e os valores dos parâmetros físicos utilizados no modelo, são apresentados em Johannesmeyer et al. (2002).

A reação irreversível clássica de primeira ordem ($A \rightarrow B$) é considerada. Os reagentes são perfeitamente misturados e os parâmetros físicos permanecem constantes ao longo do tempo. É comum que durante a modelagem do reator CSTR considere-se que este é perfeitamente misturado, ou seja, os valores das variáveis consideradas no processo, como

temperatura do reator e concentração de reagentes, por exemplo, não variam dentro do reator.

Dessa forma, as variáveis medidas na saída do reator também são iguais no interior do tanque (FOGLER, 2002). O modelo compreende o balanço de massa, energia e dos componentes descrito por um sistema de quatro equações diferenciais oridinárias (Eq. 4.1-4.4). A Figura 1 é o modelo esquemático o reator CSTR, juntamente com suas malhas de controle, as condições operacionais e os parâmetros estão apresentados no Quadro 2. Na Tabela 1 são listadas as variáveis de processo consideradas no modelo fenomenológico.

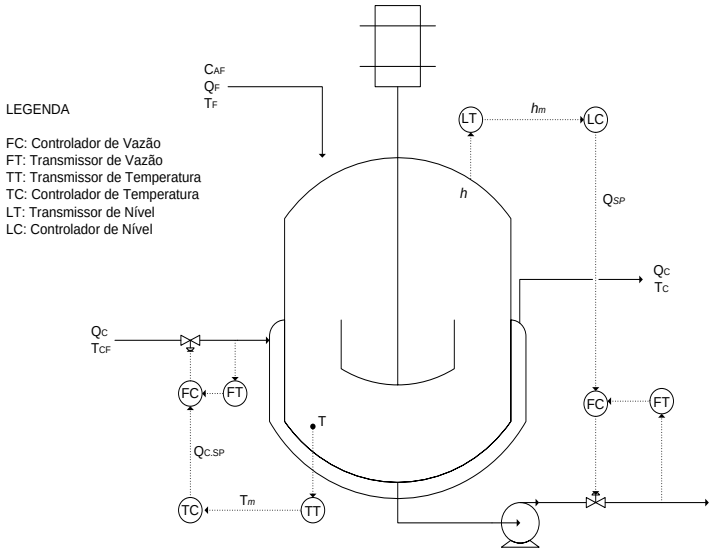
$$\frac{dC_A}{dt} = -k_0 e^{-E/RT} C_A + \frac{Q_F C_{AF} - Q C_A}{Ah} \tag{4.1}$$

$$\frac{dT}{dt} = \frac{k_0 e^{-E/RT} C_A (-\Delta H)}{\rho C_p} + \frac{Q_F T_F - QT}{Ah} + \frac{UA_C (T_C - T)}{\rho C_p Ah} \tag{4.2}$$

$$\frac{dT_c}{dt} = \frac{Q_c (T_{CF} - T_C)}{V_C} + \frac{UA_C (T - T_C)}{\rho_C C_p Ah} \tag{4.3}$$

$$\frac{dh}{dt} = \frac{Q_F - Q}{A} \tag{4.4}$$

Figura 1 – Reator de Tanque Agitado Contínuo não-isotérmico (CSTR)



Fonte: Singhal e Seborg (2002).

Quadro 2 – Variáveis do processo

Símbolos	Descrição
CAF	Concentração de entrada do reator
QF	Vazão de alimentação no reator
TF	Temperatura do reagente
QC	Vazão de alimentação da camisa
TCF	Temperatura do fluido refrigerante da camisa
T	Temperatura do reagente dentro do reator
Q	Vazão de saída do reator
Tc	Temperatura de saída da camisa
CA	Concentração de saída da mistura

Fonte: Singhal e Seborg (2002).

Tabela 1 – Condições operacionais nominais e parâmetros do modelo

Símbolo	Descrição	Valor	Unidade
A	Área da seção transversal do reator	0,1666	m^2
C_A	Concentração da espécie A no reator	0,037	mol/L
C_{AF}	Concentração da espécie A na corrente de alimentação	1	mol/L
ρC_P	Capacidade de líquido refrigerante	239	J/(L.K)
$\rho_C C_{PC}$	ρ_C - Capacidade calorífica da mistura C_{PC} - Capacidade calorífica do líquido refrigerante	4175	J/(L.K)
h	Nível do líquido no reator	0,6	m
ΔH	Calor de reação	$-5x10^4$	J/mol
K_0	Fator pré-exponencial da equação de Arrhenius	$7,2X10^{10}$	min^{-1}
Q	Vazão volumétrica de saída do reator	100	L/min
Q_C	Vazão volumétrica de líquido refrigerante	15	L/min
E/R	E - Energia de ativação R - Constante universal dos gases	8750	K
T	Temperatura do reator	402,35	K
T_C	Temperatura de líquido refrigerante da camisa	345,44	K
T_{CF}	Temperatura de entrada do líquido refrigerante	300	K
T_F	Temperatura da corrente de alimentação do reator	320	K
UA_C	U - Coeficiente de troca térmica A_C - Área da troca térmica	$5x10^4$	J(min . K)

Fonte: Singhal e Seborg (2002)

4.1.2 Geração do banco de dados

A proposta deste estudo de caso é diagnosticar dois tipos de operações anormais do CSTR, ambas relacionadas a distúrbios na vazão de alimentação do reator (Q_F). O conjunto total dos dados compreende 60 objetos referentes aos dois tipo de falha (Tabela 2). As falhas consideradas foram: perturbação do tipo degrau e do tipo oscilação (amortecida e sustentada). Diferentes intensidades de degrau e de frequência e amplitude de oscilações foram simulados de acordo com a Tabela 3.

Tabela 2 – Tipos de Falha

Tipo de Falha	Quant. de objetos
Degrau	30
Oscilação (amortecida e sustentada)	30

As condições para geração de cada um dos tipos de falha (degrau e oscilação) é visto na tabela 5. Em todas as simulações o estado estacionário inicial compreendeu a condição de operação normal do reator.

Tabela 3 – Condições de operação anormal

Condição Operacional	Descrição	Valor Nominal
Degrau em Q_F	mudança de passo na taxa de fluxo na vazão de entrada.	+/- 10 L/min
Oscilação amortecida em Q_F	o fluxo de alimentação muda como $e^{-t/33}sin(2\pi/10)$ L/min.	10 L/min
Oscilação sustentada de alta frequência em Q_F	oscilações sustentadas de frequência de 3 ciclos / min.	+/- 10 L/min

Fonte: Singhal e Seborg (2002).

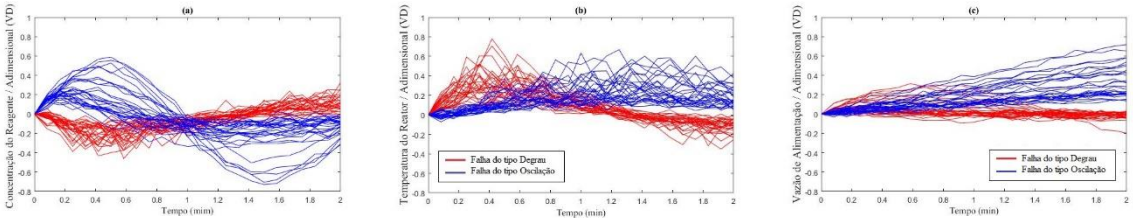
A geração das 30 séries temporais relacionadas à falha do tipo degrau considerou a amplitude do degrau no intervalo -/+10 L/min. Na falha do tipo oscilação as séries temporais foram submetidas a parâmetros de frequência da oscilação, a amplitude da oscilação e o argumento γ de uma potência do tipo, $e^{\gamma\frac{-t}{33}}$, que constitui uma função temporal da amplitude, decrescente no tempo. Para oscilação amortecida o parâmetro λ é um número real positivo, enquanto para oscilação sustentada esse argumento tem valor nulo (Eq. 4.5).

$$Q = e^{\gamma\frac{-t}{33}}sen(2\pi/10) * Q_{nominal} \quad (4.5)$$

onde $Q_{nominal}$ é o valor da vazão na condição nominal de operação do CSTR ($Q_{nominal}$ = 100L/min).

A Figura 2 apresenta as séries temporais associadas às seguintes variáveis: concentração do reagente (C_A), temperatura do reator (T) e vazão na alimentação do reator (Q_F). A janela do período de amostragem foi de 2 minutos para todos os objetos. Cada série temporal foi normalizada considerando os valores máximos e mínimos de cada uma das variáveis em toda a amostra (60 objetos) e, em seguida, cada série foi representada na forma de Variavel Desvio (VD), que consiste em subtrair todos os pontos de seu respectivo valor inicial.

Figura 2 – Séries temporais multivariadas (CSTR): (a) concentração de reagente no reator; (b) temperatura do reator; (c) vazão de alimentação do reator.

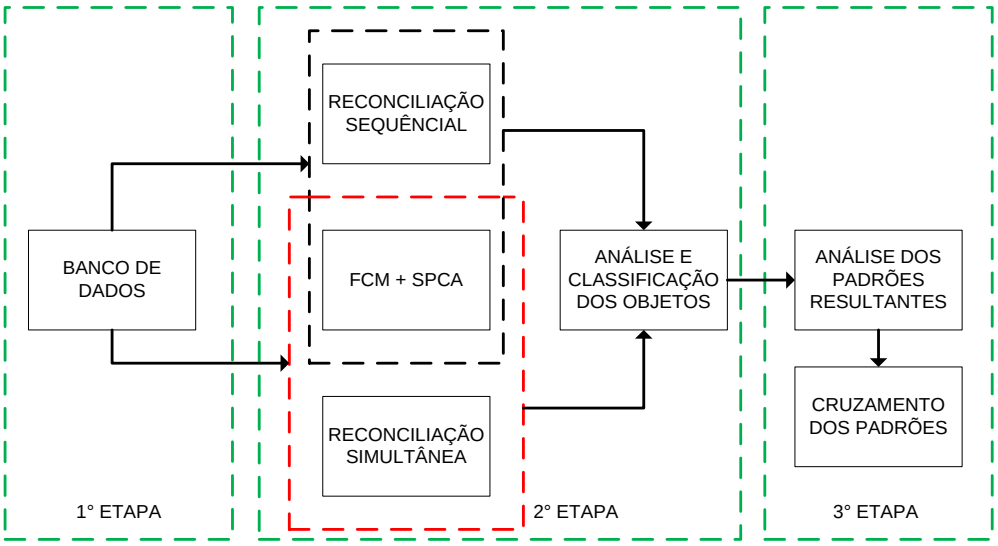


O banco de dados gerado pelo reator foi submetido às duas propostas de reconciliação (sequencial e simultâneo) e ao método do algoritmo FCM com o objetivo de comparar os padrões resultantes. A primeira abordagem contempla o caso de reconciliação sequencial (Eq. 3.2), que considera apenas variáveis de saída (T e C_A). Na segunda abordagem, a reconciliação simultânea (Eq. 3.9) foi aplicada, além das variáveis de saída, a vazão de alimentação (Q_F) (variável de entrada) também foi incluída no conjunto de dados.

Considerando que em ambos os casos o treinamento compreendeu um aprendizado não supervisionado (dados não rotulados previamente), os resultados de classificação foram utilizados como parâmetro para avaliar a qualidade do agrupamento obtido, ou seja, é um requisito importante que os padrões reconciliados sejam capazes de manter, ou mesmo melhorar, a qualidade dos resultados de classificação. Por sua vez, o cruzamento entre os padrões reconciliados e não reconciliados consistiu na verificação da consistência dinâmica dos padrões reconciliados. Comparativamente, há possíveis inconsistências ou distorções nos padrões não reconciliados.

As etapas realizadas na condução desse primeiro estudo de caso são apresentadas na Figura 3.

Figura 3 – Metodologia proposta no 1° estudo de caso



4.2.3 Resultados

O problema de otimização é resolvido pelo método clássico de segunda ordem (função *fmincon*) do *software* MATLAB, sendo este um método númerico, uma vez que a resolução análitica não se aplica.

Considerando um amostra total de 60 objetos (não rotulados) o método FCM e o método da reconciliação sequencial foram aplicados apenas nas variaveis de saída do reator C_A e T (Figuras 2a e 2b). O critério de validação do agrupamento compreendeu os resultados de classificação dos objetos. A Tabela 4 apresenta os percentuais de classificações incorretas dos padrões não reconciliados (apenas FCM) e com padrões reconciliados (FCM e posterior reconciliação).

Tabela 4 – Porcentagem de erro de classificação: FCM e reconciliação sequencial (CSTR)

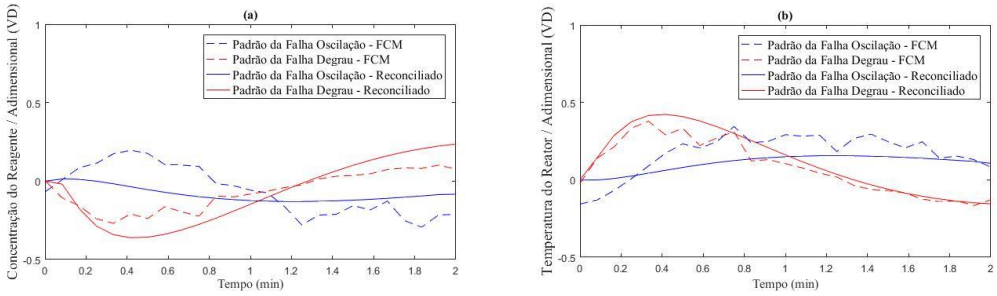
Métodos / **Tipo de Falhas**	**Padrões não Reconciliação (FCM)**	**Padrões Reconciliação Sequencial**
Degrau	0%	10% (3 objetos)
Oscilação	17% (5 objetos)	3% (1 objeto)

Diante do resultados da Tabela 4, mesmo o método FCM com 5 objetos má classificados não comprometeu o agrupamento, isso evidencia a robustez do algoritimo FCM

e a capacidade do índice SPCA em dignosticar com bastante eficiênca o tipo de falha (distúrbios degrau e oscilação na vazão de alimentação), apenas reconhecendo as diferenças no comportamento dinâmico das variáveis de saídas. Assim como os resultados da reconciliação sequencial que não comprometeram a qualidade do agrupamento e da classificação.

A Figura 4 mostra o cruzamento dos padrões (reconciliados e não reconciliados) das falhas ou pertubação do tipo degrau e osciliação (sustentada e amortecida).

Figura 4 – Cruzamento dos Padrões: FCM e Reconciliação Sequencial: (a) concentração do reagente; (b) temperatura do reator



Conforme a Figura 4, os padrões (reconciliado e não reconciliado) associados à falha do tipo degrau em ambas as variáveis de saída são similares do ponto de vista dinâmico, o que se justifica pela homogeneidade das séries temporais relacionadas a este tipo de falha (Figuras 2a e 2b). Por outro lado, é possível constatar uma inconsistência na dinâmica dos padrões não reconciliados (falha do tipo oscilação), visto que o aumento inicial da temperatura do reator (Figura 4b), ocasionado pela perturbação em Q_F, deveria provocar uma redução na concentração do reagente (o que não se verifica, Figura 4a). Além disso, a redução subsequente na concentração do reagente (em torno de 0,5 minutos) não seria esperada visto que não houve um aumento significativo na temperatura do reator. Ou seja, há um desacoplamento dinâmico/fenomenológico entre os padrões de temperatura e concentração não reconciliados para a falha do tipo oscilação.

Ambos os padrões reconciliados associados à falha do tipo oscilação apresentaram comportamento dinâmico conjugado consistente com o processo, mostrando que a reconciliação foi capaz de excluir efeitos de ruído senoidal intencionalmente inserido nas séries temporais de concentração de reagente (Figura 2a) e, ao mesmo tempo, manteve a qualidade da partição. Por outro lado, os padrões resultantes do algoritmo FCM (sem

reconciliação), mostram a incapacidade de garantir a coerência dos centros reconhecidos, nos quais os dados apresentam apenas uma pequena heterogeneidade entre os objetos da mesma classe (mesmo tipo de falha) que é muito comum em dados reais.

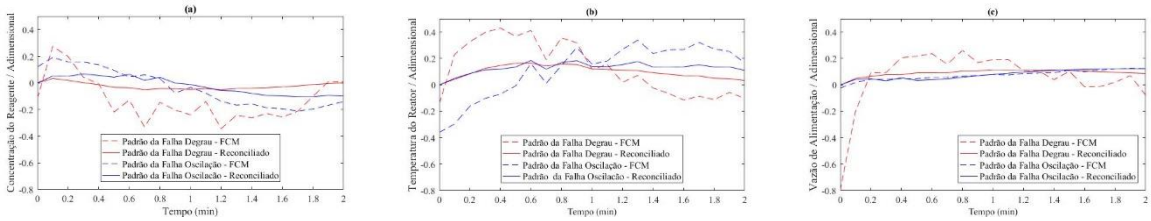
A segunda abordagem envolveu o FCM e a reconciliação simultânea (Eq. 3.10). Neste caso, além das variáveis de saída (C_A e T), o problema envolveu também variável de entrada (Q_F), considerando-se as mesmas falha do tipo degrau e oscilação (amortecida e sustentadas). Os resultados de classificação são apresentados na Tabela 5, avaliando ambos os dois métodos (reconciliados e não reconciliados) para a falha degrau os percentuais de classificações incorretas foi de apenas três objetos.

Tabela 5 – Porcentagem de erro de classificação: FCM e reconciliação simultânea (CSTR)

Métodos / **Tipo de Falhas**	**Padrões não Reconciliação (FCM)**	**Padrões Reconciliados Simultâneamente**
Degrau	10% (3 objetos)	10% (3 objetos)
Oscilação	0%	0%

No caso da falha do tipo oscilação (amortecida e sustentada), em ambas as abordagens (com e sem reconciliação) os resultados de classificação foram melhores aos da Tabela 4, mostrando que a inclusão da vazão de alimentação (Q_F) no problema de otimização agrega informações para o aprendizado não supervisionado. Os padrões resultantes estão apresentados na Figura 5.

Figura 5 – Cruzamento dos Padrões: FCM e Reconciliação Simultânea. (a) concentração do reagente; (b) temperatura do reator; (c) vazão de alimentação.



Apartir dos padrões obtidos em ambos os métodos (reconciliados e não reconciliados) apresentados na Figura 5a (vazão de alimentação) é proposto uma análise de viabilidades desses padrões que são apresentados nas Figura 6 e Figura 7, ou seja, através desta variavel de entrada é possível verificar se os padrões de temperatura e concentração (variaveis de saída)

reconhecidos em cada grupo estão consistentes dinamicamente com o efeito provocado pela vazão de alimentação segundo o modelo do processo. A Figura 6 representa o cruzamento dos padrões resultantes reconciliados simultaneamente e os padrões preditos pelo modelo, e a Figura 7 é o cruzamento dos padrões obtidos pelo FCM e o preditos pelo modelo.

Figura 6 – Teste de Viabilidade de Padrões: Modelo e Reconciliação Simultâneo: (a) Concentração do reagente; (b) Temperatura do Reator.

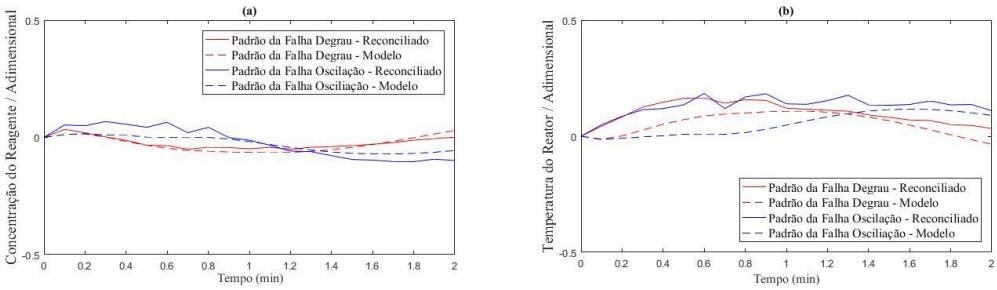
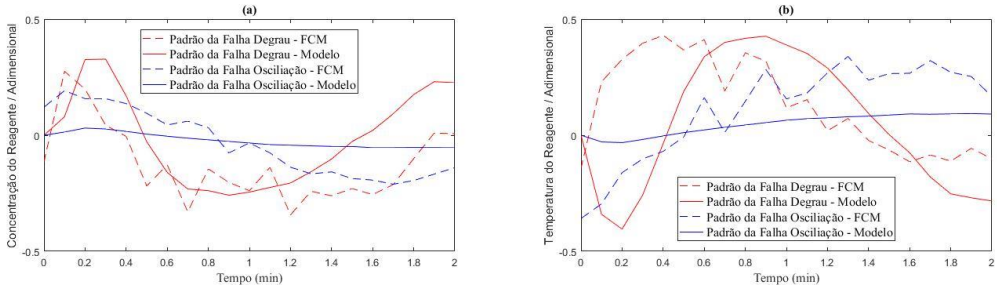


Figura 7 – Teste de Viabilidade os Padrões: Modelo e FCM e fator SPCA Concentração do Reagente; (b) Temperatura do Reator.



Os padrões resultantes dos métodos não reconciliados (FCM) e reconciliados, cruzados com os seus respectivos padrões preditos pelo modelo das variáveis concentração e de temperatura referentes à falha do tipo oscilação (Figura 6a e Figura 7a) têm comportamento dinâmico similares, o que se justifica pelo fato da entrada (Q_F) em ambos os casos (Figura 5c) são bastantes similares. Porém, na falha do tipo degrau os padrões reconciliados apresentam um comportamento dinâmico para C_A e T muito próximo do modelo, diferente dos padrões da obtidos pelo FCM que apresenta uma dinâmica dissimilar, mostrando a incosistência deste último em relação à vazão de entrada. Desta maneira, conclui-se que os padrões não reconciliados para falha do tipo degrau não são factíveis.

4.2 ESTUDO DE CASO 2: DETECÇÃO DE FALHAS EM UMA TURBINA A GÁS

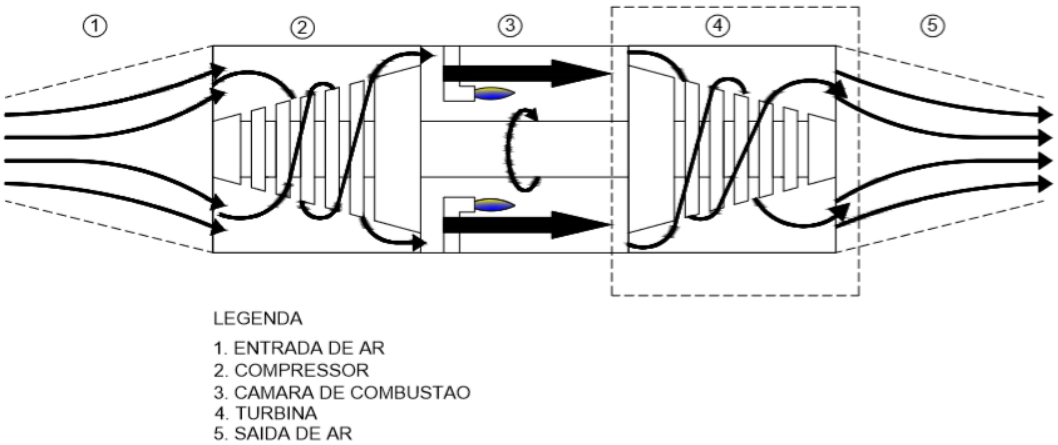
4.2.1 Descrição do Processo

O segundo estudo de caso compreendeu dados históricos de operação em uma Unidade Termoelétrica composta por três turbinas a gás com a capacidade de geração de 27MW, cada uma. Esta Unidade (UTE Rômulo Almeida, Camaçari-Ba) é parte integrante do parque da Petrobras. O principal insumo utilizado na UTE é o gás natural, com uma capacidade de processamento de 260.3 t\h de vapor e uma geração de 137 MW de energia elétrica (BARRAGAN et al., 2012; FONTES, PEREIRA, 2016).

Uma turbina a gás (Figura 8) é composta por um compressor, uma câmera de combustão e a turbina propriamente dita. Esse conjunto de equipamentos opera em um ciclo aberto, ou seja, o ar atmosférico é admitido para em seguida ser comprimido, resultando em uma maior quantidade de ar na câmara de combustão, o que significa em uma queima mais rica, garantindo maior potência do equipamento.

Desta forma, o ar comprimido é então pulverizado com combustível e uma faísca elétrica acende a mistura. Os gases de queima se expandem rapidamente e são esgotados através da parte traseira da câmara de combustão. Estes gases exercem a mesma força em todas as direções, como mostra a Figura 8, proporcionando um impulso de avanço enquanto eles escapam para a turbina fazendo girar o seu eixo.

Figura 8 – Turbina a gás

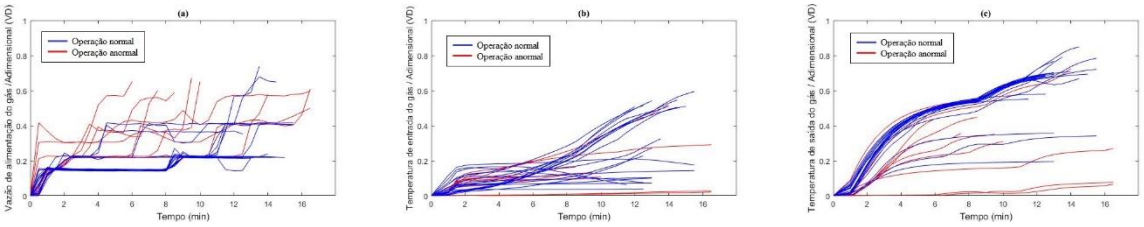


A energia produzida pela turbina da UTE é fornecida ao sistema interligado nacional, sendo controlada e monitorada pelo Operador Nacional do Sistema Elétrico (ONS). De acordo com o regulamento da ONS, havendo qualquer anormalidade no fornecimento de energia elétrica e no fornecimento de vapores de alta e baixa pressão, a Unidade Termoelétrica fica sujeita a penalidades e multas contratuais.

Diante ao risco do não cumprimento do fornecimento de tais insumos, a gestão operacional da UTE deve prever possíveis falhas que acarretam na interrupção do processo. Na UTE Rômulo Almeida, o modelo das turbinas a gás utilizadas é o RB211-G62DF, do fabricante Rolls-Royce (FONTES; PEREIRA, 2016). Por questões de segurança, existe um sistema de controle para desarmar o equipamento caso a temperatura de alguns dos sensores localizados na câmera de combustão ultrapasse os valores médio dos demais em +/- 150°C. Este tipo de ocorrência denomina-se "*trip*" por dispersão de temperatura.

A Figura 9 representa todas as séries temporais da amostra de treinamento, obtidas diretamente do histórico do processo. O período de amostragem adotado foi de aproximadamente de 17 minutos e, tal como no primeiro estudo de caso, cada série temporal foi normalizada no intervalo [0;1], considerando os valores máximo e mínimo de cada variável em toda a amostra. Em seguida, cada série foi representada na forma de VD. Conforme se verifica, as janelas de tempo não possuem o mesmo tamanho uma vez que o tempo de partida da turbina não é necessariamente o mesmo em todas as partidas realizadas. Portanto, o SPCA foi aplicado para mensurar a similaridade entre séries multivariadas associadas a diferentes janelas de tempo.

Figura 9 – Séries Temporais Multivariada (Turbina): (a) vazão de alimentação do gás; (b) temperatura de entrada do gás; (c) temperatura de saída do gás.



O desarme da turbina (*trip*) por dispersão de temperatura é considerado uma falha que

pode estar associada a diversas causas tais como surgência, vibração, contaminação na corrente de gás, entre outros. Segundo Fontes e Pereira (2016), não há nenhuma informação prévia de um padrão de falha desse equipamento, nem mesmo por parte do fabricante. Diante desse cenário, este estudo de caso compreendeu a aplicação do método proposto de agrupamento e reconhecimento de padrões visando identificar possíveis padrões de operações normais e anormais (falha) para o equipamento.

4.2.2 Geração do Banco de Dados

O banco de dados é composto por séries temporais (Figura 9) coletadas durante a partida da turbina no período entre 2008 e 2011 e que foram disponibilizados pelo Sistema de Gerenciamento de Informações da Planta (PIMS). A amostra total é composta por 70 objetos, sendo 60 objetos de operação normal e 10 objetos de falha (*trip*).

As variáveis de processo consideradas neste caso foram: temperatura de entrada do gás (T_E), vazão de alimentação de gás (Q_A) e temperatura de saída (T_S). Portanto, duas entradas e uma única saída. Os objetos possuem, entretanto, diferentes comprimentos de janelas de tempo e a maior janela de tempo entre todos os objetos da amostra foi de 16 minutos.

O conjunto total de dados foi dividida em amostras de treinamento e teste (Tabela 6). A amostra de treinamento é composta por um total de 40 objetos, sendo 10 objetos de operação com falha e 30 objetos de operação normal. A amostra de teste compreendeu 30 objetos de operação normal, aleatoriamente selecionados entre os 60 objetos da amostra original associados à partida normal.

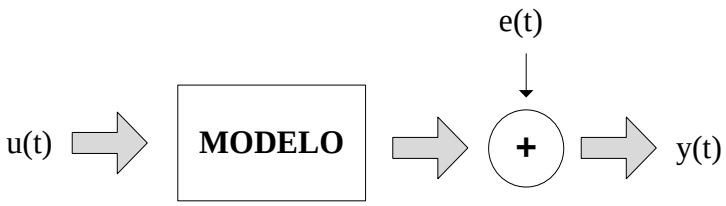
Tabela 6 – Amostra de dados: treinamento e teste

Amostra de objetos	Objetos Normais	Objetos com Falha	Total
Amostra de treinamento	30	10	40
Amostra de teste	30	0	30

4.2.3 Modelo Auto-regressivo com Entradas Externas (ARX)

A estrutura de modelo auto-regressivo com entradas externas (ARX, *autoregressive with exogenous inputs*) compreende uma abordagem dinâmica empírica (Figura 10) devidamente consolidada na área de identificação de sistemas (AGUIRRE, 2004).

Figura 10 – Estrutura do Modelo ARX



A estrutura de modelo ARX para o caso de uma entrada e uma saída (SISO, *Single Input, Single Output*) é dada pela Eq. 4.6.

$$y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) = b_1 u(t-n_k) + \cdots + b_{n_b} u(t-n_b-n_k+1) + e(t) \quad (4.6)$$

onde *y(t)* é a variável de saída no tempo *t*, *u(t)* é a variável de entrada no tempo *t*, n_a é o número de termos passados da saída, n_b é o número de termos passados da entrada e n_k é o tempo morto da variável de entrada *u* associado à saída *y*. *e(t)* representa ruído branco (AGUIRRE, 2004).

A estrutura ARX pode também ser representada na seguinte forma compacta (Eq. 4.7):

$$A(q)y(t) = B(q)u(t-n_k) + e(t) \quad (4.7)$$

Onde:

$$A(q) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a} \quad (4.8)$$

$$B(q) = b_1 + b_2 q^{-1} + \cdots + b_{n_b} q^{-n_b+1} \quad (4.9)$$

q é um operador deslocamento $\left(u(t) \cdot q^{-z} = u(t-z)\right)$.

O modelo ARX foi escolhido neste trabalho para representar a turbina, pelo fato de não haver um modelo fenomenológico capaz de descrever o seu comportamento dinâmico.

O conjunto de dados da turbina envolve duas variáveis de entrada (vazão e temperatura de admissão do gás natural, F_g e T_i) e apenas uma variável de saída (temperatura de saída do gás, T_e). Desta forma, a estrutura ARX utilizada para representar a turbina compreendeu o seguinte modelo (Eq. 4.10):

$$\hat{T}_e(k_t) - 0.699 \cdot \hat{T}_e(k_t - 1) = -0.266 \cdot F_g(k_t) + 0.487 \cdot$$
$$F_g(k_t - 1) + 1.978 \cdot T_i(k_t) - 1.671 \cdot T_i(k_t - 1) + e(k_t) \quad (4.10)$$

$\hat{T}_e(k_t)$ é a saída predita no instante de tempo k_t. Este modelo descreve o efeito dinâmico das entradas (F_g e T_i) e sobre a saída (T_e).

Todos os parâmetros, tempo morto e número de valores passados de cada uma das entradas e da saída (ordens do modelo) foram selecionados a partir de um conjunto de opções previamente estabelecido com o objetivo de obter o melhor ajuste em relação às medições disponíveis para a variável de saída (T_e). Por outro lado, cada objeto compreende uma série temporal multivariada (STM) com três variáveis de acordo com a Figura 9, com a mesma janela de tempo, porém os objetos estão relacionados a diferentes períodos de operação. Portanto, a estrutura de dados consiste em uma matriz tripla típica envolvendo objetos ("lotes") x variáveis x tempo.

A soma global dos erros de previsão em toda a amostra foi usada como métrica para avaliar a qualidade do modelo (Eq. 4.11).

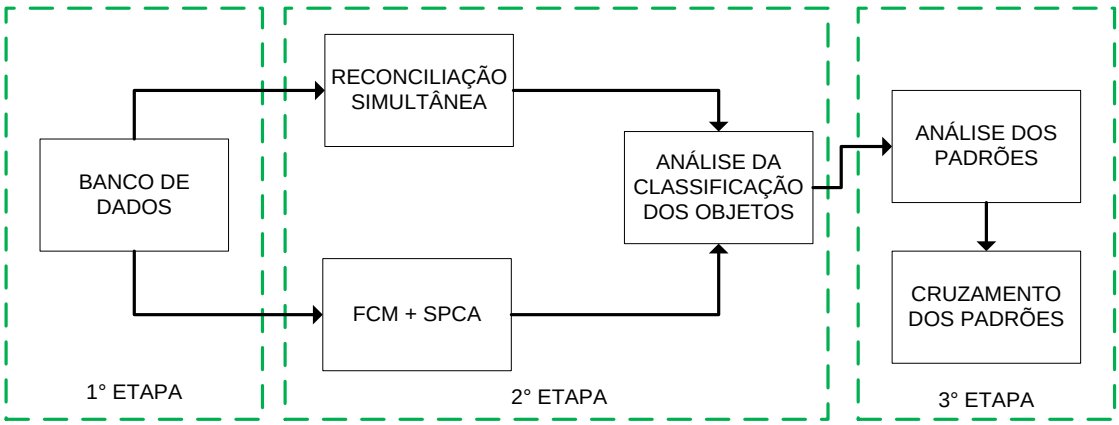
$$\sum_{i=1}^{70} \sum_{j=1}^{m_i} \left(\hat{T}_e(k_{tj}) - T_{ei}(k_{tj}) \right)^2 \quad (4.11)$$

m_i é a duração da janela de tempo do objeto *i* (i = 1, ..., 70) e $\hat{T}_e(k_{tj})$ e $T_{ei}(k_{tj})$ são os valores da saída predita e medida do objeto *i* no instante do tempo k_{tj}.

A abordagem deste estudo de caso contempla o método FCM e apenas a reconciliação simultânea (Eq. 3.9) utilizando o ARX identificado (Eq. 4.6). Após a análise dos resultados de classificação, os padrões reconhecidos (reconciliados e não reconciliados) foram cruzados para a identificação de possíveis inconsistências.

As etapas realizadas na condução desse primeiro estudo de caso são apresentadas na Figura 11.

Figura 11 – Metodologia proposta no 2º Estudo de caso



4.2.4 Resultados

Neste estudo de caso, o problema de otimização é resolvido com o mesmo método numérico do anterior. Os resultados têm como base as amostras de treinamento e de teste (Tabela 6). A primeira análise foi feita a partir dos resultados de classificação de ambos os métodos (reconciliado e não reconciliado). Os percentuais de classificação incorretas nas amostras de teste e treinamento são apresentados na Tabela 7. Os dados referentes ao grupo de falha, mesmo sendo uma amostra minoritária, nas duas abordagens (com e sem reconciliação) foram capazes de reunir 80% dos objetos em um mesmo grupo. Os trabalhos de Fontes e Pereira (2016) e Fontes e Budman (2017), relacionados com essa mesma amostra de dados, mostram que 20% (2 objetos) de classificações incorretas relacionadas à partida com falha foram os melhores resultados obtidos a partir da amostra disponível. É possível que exista um padrão adicional de falha que não foi identificado por conta da pequena quantidade de objetos associados a esta classe. Os grupos de partida normal de ambos os métodos apresentaram 87-90% de objetos normais, respectivamente, presentes em toda a amostra de treinamento.

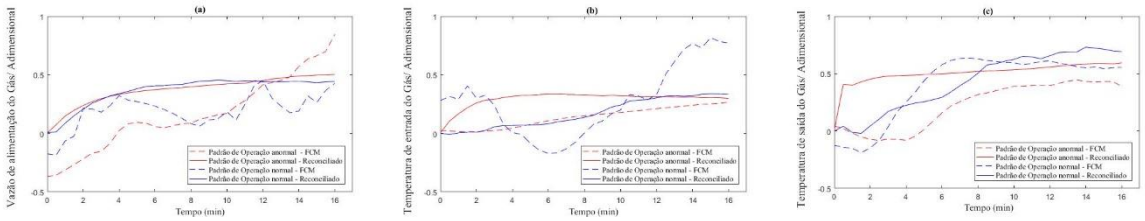
A amostra de teste (30 objetos apenas normais) reconhecida para cada grupo foram utilizados como referência para a classificação (de acordo com a similaridade) ambas as abordagens tiveram apenas 3,33% de objetos incorretamente classificados, sendo este um resultado satisfatório.

Tabela 7 – Porcentagem de erro de classificação: FCM e reconciliação simultânea (Turbina)

Métodos / Tipo de Amostra	Padrões não reconciliados		Padrões reconciliados	
Tipos Operação	Falha	Normal	Falha	Normal
Amostra de Treinamento	20%	10%	20%	13%
Amostra de Teste	-	3,3%	-	3,3%

Os padrões reconciliados e não reconciliados são apresentados na Figura 12. Nos padrões reconciliados, apesar das pequenas diferenças entre as dinâmicas do fluxo de alimentação do gás em ambos os padrões de operação normal e com falha (Figura 12a), os perfis relacionados às temperaturas de entrada do gás (Figura 12b) são menos similares, o que justifica as diferenças apresentadas na variável de saída (Figura 12c). Além disso, as diferenças entre os padrões reconciliados da operação com falha e normal são mais acentuadas no comportamento dinâmico das temperaturas de entrada e saída que são suficientes para gerar mudanças nas direções dos componentes principais capazes de reconhecer diferenças entre objetos normais e de falta. O mesmo não ocorre com o FCM, uma vez que as entradas não são bem reperesentadas na variavel de saída (Figura 12c, temperatura de saída do gás).

Figura 12 – Cruzamento dos Padrões: FCM e Reconciliação Simultâneo: (a) vazão de alimentação do gás; (b) temperatura de entrada do gás; (c) temperatura de saída do gás.

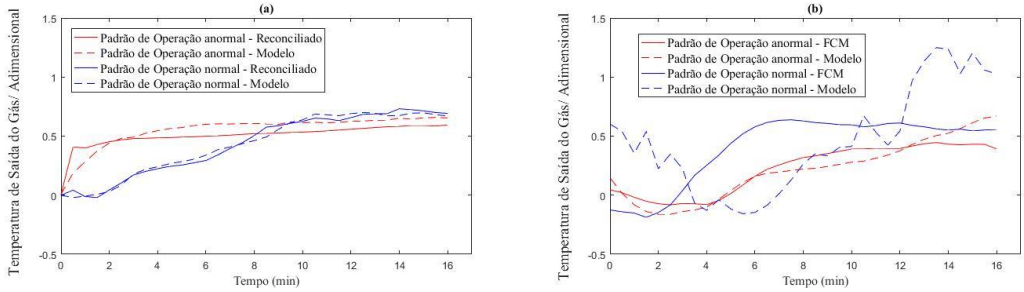


A fim de avaliar os padrões resultante de ambos os métodos, propõe-se a análise de viabilidade, partindo do mesmo principio do estudo de caso anterior. A Figura 13 apresenta os padrões de saída (temperatura de saída do gás) preditos pelo modelo ARX considerando os mesmos perfis de entrada (temperatura e vazão de entrada do gás, 12a-b, respectivamente). No cruzamentos dos padrões apresentados na Figura 13, ressalta-se que a Figura 13a corresponde ao padrão predito pelo modelo reconciliado e a dinâmica destes estão bem próximas, comprovando consistências dos padões de entrada. Ao contrário dos padrões

resultantes do FCM (Figura 13b) tem-se diferenças significativas entre os padrões não reconciliados da temperatura de saída do gás e a dinâmica dominante predita para o processo.

À luz disso, pode-se dizer que o FCM (modelo não reconciliado), a depender da complexibilidade do problema de otimização, é incapaz de obter padrões coerentes com o modelo do processo, diferente dos padrões reconciliado., tendo em vista que existe uma relação entre as variaveis de entrada e as variaveis de saída com o modelo o processo.

Figura 13 – Teste de Viabilidade (Turbina a Gás). (a) Reconciliação e Modelo; (b) FCM e Modelo



CAPÍTULO 5

Conclusão e sugestões para trabalhos futuros

De forma inovadora, um método de otimização para o reconhecimento de padrões em séries temporais foi desenvolvido e aplicado com êxito em dois estudos de caso, um simulado e outro real. Ao longo deste trabalho, o problema de análise de agrupamento usando o algoritmo *fuzzy c-means* (FCM) foi formulado e resolvido, levando à identificação de um algoritmo reconciliado para cenários indústrias.

Nesta dissertação, foi feita uma revisão bibliográfica sobre análise de agrupamento, assim como a investigação de métrica de similaridade para séries temporais multivariados. A partir deste levantamento bibliográfico e testes realizados, identificou-se inconsistência no algoritmo FCM. Desta forma, apresentou-se uma nova metodologia, na qual a estratégica é baseada no algoritmo FCM clássico, que considera a dinâmica do processo com restrição para garantir a viabilidade dos padrões reconhecidos.

A proposta geral dessa nova metodologia consiste em uma otimização bi-critério sujeita a restrições severas associadas às variáveis de processo, modelo e graus de pertinência dos objetos aos grupos. Portanto, este trabalho apresenta uma alternativa viável para incluir conhecimentos prévios relacionados ao comportamento dinâmico do processo para guiar um algoritmo de agrupamento e reconhecimento de padrões envolvendo séries temporais multivariadas.

Segundo o método mencionado, o algoritmo desenvolvido foi aprimorado em duas estratégica de reconciliação (sequencial e simultânea) e adaptado ao problema de diagnóstico de falhas em processos indústrias. O primeiro estudo de caso compreendeu uma unidade virtual com Reator Tanque Agitado Contínuo (CSTR), usada como referência para estudos de detecção de falhas e controle. A planta virtual do Reator Tanque Agitado Contínuo (SEBORG, 2005), foi implementada e simulada com êxito. O segundo estudo de caso envolveu a detecção de falha durante a partida de uma turbina a gás em uma unidade termelétrica.

Em problemas que envolvem detecção de falhas, padrões relacionados à condição normal e/ou anormais, identificados/reconhecidos através de um procedimento de agrupamento podem ser usados como referência para prever trajetórias corretas para operação normal ou para apoiar o desenvolvimento de um sistema supervisório para monitoramento em tempo real, diagnóstico e previsão de falhas. Nesses casos, é importante que os protótipos/padrões (ou referências) sejam realizáveis ou consistentes com o comportamento dinâmico dominante do processo, apresentando, entre outras características, sinais coerentes de ganhos estáticos entre entradas e saídas.

Pode-se afirmar que o objetivo geral da pesquisa de encontrar possíveis inconsistências no algoritmo FCM foi alcançado. De acordo com os testes realizados, o FCM não garantiu a coerência dos padrões em relação ao processo. Além disso, o reconhecimento de padrões viáveis seria mais difícil nos casos em que a amostra de dados tem pouca informação e há baixa homogeneidade entre objetos da mesma classe (mesmo rótulo). Esta é uma situação típica e frequente em dados extraídos de sistemas industriais reais sujeitos à ruído e perturbações desconhecidas.

Diferente das duas estratégias de reconciliação de padrões que foram capazes de obter um padrão reconciliado sem comprometer a qualidade dos resultados de agrupamento e classificação, ou seja, sem prejudicar a capacidade de diagnosticar ou detectar falhas. Portanto, a metodologia proposta fornece um modelo de classificação satisfatório de objetos a partir de padrões realizáveis.

Embora esse trabalho seja baseado no uso do SPCA como uma métrica de similaridade, a abordagem de reconciliação proposta não se limita a uma métrica específica. Outras métricas de similaridade podem ser empregadas sem alterar a essência da estratégia e conceitos propostos.

Como recomendações para trabalhos futuros, são sugeridas algumas alternativas:

a) Reformulação do problema de reconciliação envolvendo otimização multi-objetivo e comparação dos resultados com a estratégia proposta neste trabalho;

b) Adoção de outras métricas de similaridade e análise de aderência de cada métrica em relação à capacidade de obtenção de padrões factíveis;

c) Definição de uma métrica para quantificação da qualidade de um padrão de acordo com os resultados de classificação e consistência com a realidade do processo;

d) Utilização de uma abordagem de agrupamento de séries temporais multivariadas baseada em similaridade entre modelos em lugar de similaridade entre as séries originais.

5.1 PUBLICAÇÕES ACADÊMICAS

- Artigos Submetidos e Aceitos em Congresso.

  Izete Silva, Pedro Aragão, Cristiano Fontes, Raony M. Fontes, Marcelo Embiruçu. “Reconhecimento e reconciliação de padrões de series temporais em processo de produção”. XXXVII encontro nacional de engenharia de produção (ENEGEP), DOI: 10.14488/enegep2017_tn_sto_243_410_32309, Out 10-13 (2017).

  Izete Silva, Pedro Aragão, Cristiano Fontes, Raony M. Fontes, Marcelo Embiruçu. “Reconciliation of patterns in the clustering of time series". Conference: 24th ABCM International Congress of Mechanical Engineering (COBEM), DOI: 10.26678/ABCM.COBEM2017.COB17-2900, Dez 3-8 (2017).

- Artigos Submetidos em Revista

  Cristiano Fontes, Izete Silva, Marcelo Embiruçu, Pedro Aragão. “Pattern Reconciliation: A new approach involving Constrained Clustering of Time Series”, Journal: Artificial Intelligence.

REFERÊNCIAS

Abonyi, J. and, Feil, B., 2007. “Cluster Analysis for Data Mining and System Identification”. Boston Berlin”, Editora Birkhãuser.

Aguirre, L. A., 2004. “Introdução à Identificação de Sistemas: Técnicas Lineares e Não-Lineares Aplicadas a. Sistemas Reais”, 2ª ed., Editora UFMG, Belo Horizonte.

Aparajeeta, J., Nanda, P. K., and Das, N., 2016. *“*Modified possibilistic fuzzy C-means algorithms for segmentation of magnetic resonance image”, *Applied soft computing*, Vol. 41, pp 104-119.

Barragan, J. F. M., Fontes, C. H. O., Pereira, O., Sá, S. T. B., And Pacheco, L. A., 2012. “Análises de métricas de similaridades em séries temporais para reconhecimento de padrões”. *Cadernos do IME*, Série Estatística, Vol. 33, pp. 35-50.

Berget, I., Mevik, B. and Neas, T., 2007. “New modifications and applications of fuzzy C-means methodology”. *Computational Statistics & Data Analysis*, Vol. 52, p 2403-2418.

Bezdek, J. C., 1981. “Pattern Recognition with Fuzzy Objective Function Algorithms”. New York: Plenum Press.

Bezdek, J. C., Keller, J., Krisnapuram, R. and Pal, N., 2005. “Fuzzy Models and Algorithms for Pattern Recognition and Image Processing”. Springer Science Business Media, New York, 1° Edition.

Brzyski, P., Tobiasz-Adamczyk, B., and& Knurowski T., 2012. “Relevance and reliability of the GARS scale in the elderly population” Poland, *Gerontologia Polska*, vol. 20, pp. 109–117.

Chiu, S. L., 1994. “Fuzzy Model Identification Based on Cluster Estimation”. *Journal of Intelligent and Fuzzy systems*”. n. 3, p. 267–278.

Chuang, K., Tzeng, H., Chen, S., Wu, J., and Chen, T. 2006. “Fuzzy c-means clustering with spatial information for image segmentation”. *Computerized Medical Imagens and Graphics*, Vol. 30, pp. 9-15.

Culbertson, J., Guralnik, D., and Stiller, P. F., 2018. *“*Functorial hierarchical grupoing with overlaps*”s”. Discrete Applied Mathematics,* Vol. 236, pp. 108-123.

Czernyszewicz, E., 2008. “Application of main composition analysis description of the consumer quality structure of apples, Nauka”. *Technologia. Jakość*, Vvol. 57, pp;. 119–127.

D’urso, P. and Maharaj, E. A., 2012 “Wavelets-based grupoing of multivariate time series,” *Fuzzy Sets Syst.*, vol. 193, pp. 33–61.

Dao, T., Duong, K. and Vrain, C., 2017. “Constrained grupoing by constraint programming”, *Artificial Intelligence*, Vol. 244, pp. 70-94.

Deng, X. Tian, X., and Chen, S., 2013. “Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis”. *Chemom. Intell. Lab. Syst*. Vol. 127, pp. 195-209.

Döring, C. and Lesot, M., 2006. “Data analysis with fuzzy grupoing methods”. *Computational Statistics & Data Analysis*, Vol. 51, pp. 192-214.

Dunn, J. A., 1973. “Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Grupos*”. Journal of Cybernetics,* Vol. 3, pp. 32-57.

Fávero, L. P., 2009. “Análise de dados: Modelagem Multivariada para Tomada de Decisão”. Rio de Janeiro, Editora: Elsevier Editora Ltda.

Fogler, S. C., 2002. “Elementos de Engenharia das Reações Químicas”, 3ª ed., Editora LTC.

Fontes, C. H. O ande Pereira, O., 2016. “Pattern recognition in multivariate time series – A case study applied to fault detection in a gas turbine”. *Enginnering applications of artificial intelligence*, Vol. 49, p. 10-18.

Fontes, C. H. O. and Budman, H., 2017. “A Hybrid grupoing approach for multivariate time series – a case study applied to failure analysis in a gas turbine”, *ISA Transactions*, DOI: 10.1016/j.isatra.2017.09.004

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. and Tathan, R., 2009. “Análise Multivariada de Dados”**,** Porto Alegre-Brasil, Editon Bookman, 6° Edition.

Hladnik, A. 2013. “Image compression and face recognition: two image processing applications of principal component analysis”. *International Circular of Graphic Education and Research*, Vvol. 6, pp. 56–61.

Hoppner, F., Klawonn, F., Kruse, R., and Runkler, T., 1999. “Fuzzy groupo analysis - Methods for classification, data analysis and image recognition”, New York, Editora Jonh Wiley & Son, LTD.

Izakian, H., Pedrycz, W., and Jamal, I., 2015. “Fuzzy grupoing of time series data using dynamic time warping distance”, *Eng. Appl. Artif. Intell*, Vol. 39, pp. 235-244.

Jain, A. K.; Duin, R. P. W.; Mao, J. 2000. “Statistical pattern recognition: A review. Pattern Analysis and Machine Intelligence”, *IEEE Transactions*, Volv. 22, n. 1, pp. 4-37.

Jiang, X., Liu, P., Li, Z., 2014. “Data reconciliation and gross error detection for operational data in power plants*”, Energy*, Vvol. 75, pp. 14-23.

Johannesmeyer, Michael M. C., Singhal, A., and Seborg, and Dale E., 2002. “Pattern Matching in Historical Data”. *AIChE Journal*, Vol. 48, pp. 2022-2038.

Johnson, R. A., Wichern, D. W., 2007. “Applied multivariate statistical analysis”. Pearson Prentice Hall, New Jersey.

Kesemen, O., Tezel, O., Ozkul, E., 2016. *“Fuzzy c-means grupoing algorithm for directional data (FCM4DD)”,* Expert systems with applications, Vol. 58, pp 76-82.

Khediri, B., Limam, M., and Weihs, C. 2011. “Variable window adaptive Kernel Principal Component Analysis for nonlinear nonstationary process monitoring”. *Comput. Ind. Eng.*, Vol. 61, p. 437-446.

Kolasa-Wiêcek, A, 2012. “Regression modeling of agriculture greenhouse gases emissions in Poland”. *Ecol Chem Eng A*, Vol. 19, pp. 1383-1391.

Krishnapuram, R., Keller J. M. 1993. “A Possibilistic Approach to Grupoing”. *IEEE Transactions on fuzzy systems*, Vol. 1.

Lattin, J., Carroll, J., Douglas e Green, P. E., 2011. “Análise de dados multivariados”. Cengage Learning, São Paulo, 475.

Li, S. and Wen, J., 2014. “Application of pattern matching method for detecting faults in air handling unit system,” *Autom. Constr.*, Vol. 43, p. 49–58.

Li, X., Wang, N., Wang, L., Kantor, I., Robineau, J., Yanga, Y, Maréchalb, F., 2018. *“A data-driven model for the air-cooling condenser of thermal power plants based on data reconciliation and support vector regression”*, Applied Thermal Engineering, vol. 129, pp. 1496–1507.

Liao, T. W., 2005. “Grupoing of time series data - a survey”. *Pattern Recognit*, Vol 38, p. 1857- 1874.

Memon, K. H., Lee, D., 2018. *“*Generalised kernel weighted fuzzy c-means grupoing algorithm with local information*”. Fuzzy sets and systems*, Vol. 340, pp 91-108.

Mendel, J. M., 2001. “Uncertain rule-basead fuzzy logic systems: introduction and new directions”. United States of America, Editora: Prentice-Hall.

Mingoti, S. A., 2005. “Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada”. Belo Horizonte, Editora UFMG.

Nowicki, J., Żylińska, A., Kin, A. 2013. “Application of statistical and graphic methods in the analysis of tectonically deformed trilobites from the Ellipsocephalidae Matthew family, 1887 from the Cambrian Świętokrzyskie Mountains”. XXII Scientific Conference of the Paleontological Section of the Polish Geological Society "Updateism and anti-aging in paleontology" pp. 38-39.

O’Reilly, C., Moessner, K. and Nati, Michele, 2017. “Univariate and Multivariate Time Series Manifold Learning”. *Knowledge-Based Systems*, Vol. 133, p 1-16.

Raskin, R and Terry, H., 1988. “A Principal-Components Analysis of the Narcissistic Personality Inventory and Further Evidence of Its Construct Validity”. Journal of Personality and Social Psychology, Vol. 54, pp. 890-902.

Reis, E., 2001. “Estatística Multivariada Aplicada”. Lisboa, Edições Spilabo, Lda.

Rymuza,K., Turska, E., Wielogórska, G. and Bombik, A., 2013. “Use of principal component analysis for the assessment of spring wheat characteristics”. *Acta Sci. Pol.*, Vol. 11, pp. 79-90.

Saravanamutto, H. I. H.; Rogers, G. F. C.; Cohen, H., 1996. “Gas Turbine Theory”. 5ª ed., Dorchester: Prentice Hall.

Schuch, R., Dill, S., Suasen, P., Padoin, E., and Campos, M., 2010. “Mineração de dados em uma subestação de energia elétrica”. 9 th Brazilian Conference on Dynamics, Control and their Applications.

Shuanghua, B., McLean, D. D. and Thibault, J., 2007. “Impact of model structure on the performance of dinamic data reconciliation. *Computers and Chemical Engineering*, Vol. 31, p. 127-135.

Singhal, A. and Seborg, D. E., 2002. “Pattern matching in multivariate time series databases using a moving-window approach”. *Ind. Eng. Chem. Res.* Vol. 41, p. 3822-3838.

Singhal, A. and Seborg, D. E., 2006. “Evaluation of a pattern matching method for the Tennessee Eastman challenge process”. *J. Process Control*, Vol. 16, p. 601-613.

Sorsa, T.; and Koivo, H. N., 1993. “Application of artificial neural networks in process fault diagnosis”. Automatica, n 4, p. 843-849.

Suchacz, B. and, Wesolowski, M., 2010. “Relationships between zinc, copper, lead and nickel content in wild herbal extracts”, *Bromat. Chem. Toksykol*, Vvol. 4, pp. 485-492.

Syed, M.S., Dooley, K. M., Madron, F., Knopf, F. C., 2016. *“Enhanced turbine monitoring using emissions measurements and data reconciliation”,* Vvol. 173, pp. 355-365.

Tak-chung Fu. 2011, “A review on time series data mining”. *Engineering Applications of Artificial Intelligence*. Vol. 24, p.164-181.

Trebuňa, P. and Halčinová, J., 2013. “Mathematical Tools of Grupo Analysis”. *Applied Mathematics*, Vol. 4, p. 814-816.

Vaidyanathan, R. and Venkatasubramanian, V., 1992. “Representing and diagnosing dynamic process data using neural networks”. *Engineering Applications of Artificial Intelligence*, Vol. 5 p. 11-21.

Yang, K. and Shahabi, C., 2004. “A PCA-based Similarity Measure for Multivariate Time Series”, In p*roceedings of the International Workshop on Multimedia Databases, ACM-MMDB, Washington DC, USA*, p. 1–10.

Zadeh, L. A., 1965. “Fuzzy Sets”, *Inf. and Control*, Vol. 8, p. 338-353.

Zadeh, L. A., 1999. “Fuzzy sets as a basis for a theory of possibility”. Fuzzy Sets and Systems, Vol. , pp 3-28.

Zhang, Y, Wang, Z., and J. Ma, 2011. “Fault localization in electrical power systems: A pattern recognition approach,” *Electr. Power Energy Syst.*, Vvol. 33, pp. 791–798.

ANEXO A

Pattern Reconciliation: A new approach involving Constrained Clustering of Time Series

Cristiano Hora Fontes[1*], Izete Celestina Santos[2], Marcelo Embiruçu[3], Pedro Aragão[4]

Programa de Engenharia Industrial (Graduate Program in Industrial Engineering), Escola Politécnica (Polytechnic Institute), Universidade Federal da Bahia (Federal University of Bahia, Brazil)

1 – cfontes@ufba.br, 2 – izabox@gmail.com, 3 – embirucu@ufba.br, 4 – pedrom.aragao@outlook.com

*Corresponding Author: cfontes@ufba.br

Abstract

In spite of the advances in strategies involving clustering and pattern recognition in time series, there are no approaches capable of directly associating the recognized patterns with the dynamic behavior of the process investigated. Works related to the bi-criterion constrained clustering have failed to present a systematic way of coping with the problem of clustering time series and the need to obtain feasible patterns, reconciled with the reality of the process. This paper presents a new approach involving pattern reconciliation in the clustering of time series, starting from the analysis of the simplest case (univariate time series) and proposes a generic optimization problem for the clustering of multivariate time series. The strategy is based on Fuzzy C-Means (FCM) and directly considers the process dynamics as a soft constraint in order to ensure the feasibility of the recognized patterns. The proposed method is applied in two case studies. The first comprises the diagnosis of abnormal (failures) operation of a non-isothermal Continuous Stirred Tank Reactor (CSTR), a well-known benchmark system used for the assessment of Fault Detection and Diagnosis (FDD) techniques. The second comprises a real industrial scenario which involves the recognition of starting patterns in a gas turbine for fault detection purposes. The results show that the proposed method for reconciling patterns (FCM coupled with the process model) is able to recognize feasible patterns preserving the quality of clustering and classification.

Keywords: process model; pattern reconciliation; time series; constrained clustering

1. Introduction

Time series are an important class of temporal data widely used in industrial processes in which the consolidation of data acquisition techniques has been encouraged by the development of Data Mining (DM) methods for the extraction of process knowledge. One of the applications of time series extracted from the databases comprises clustering and pattern recognition ((Liao, 2005); (Kavitha and Punithavalli, 2010); (Aghabozorgi, Shirkhorshid and

Wah, 2015); (Fu, 2011); (Izakian, Pedrycz and Jamal, 2015)) These approaches can provide a feasible and useful alternative way of detecting and/or diagnosing abnormal (failures) operation ((Venkatasubramanian *et al.*, 2003); (Fontes and Pereira, 2016); (Fontes and Budman, 2017)).

Data clustering and pattern recognition in time series have been investigated through the use of traditional techniques, especially in the case of univariate series ((Liao, 2005); (Keogh and Kasetty, 2002); (Trebuňa and Halčinová, 2013), (Zakaria *et al.*, 2016)). This kind of problem can be solved using point-prototype clustering models such as the traditional Fuzzy C-Means (FCM) algorithm ((Bezdek *et al.*, 2005)) based on standard similarity metrics (Euclidean and Dynamic Time Warping, DTW) distances ((Wang *et al.*, 2013); (Bankó and Abonyi, 2012); (Petitjean, Ketterlin and Gancarski, 2011)). The clustering of Multivariate Time Series (MTS) represents a more complex problem due to its intrinsic features such as the choice of the similarity metric and clustering validation. One of the widely used similarity metrics for the comparison of two multivariate time series comprises PCA-based similarity metrics (SPCA) and its different versions ((Singhal and Seborg, 2006); (Khediri, Limam and Weihs, 2011); (Deng, Tian and Chen, 2013); (Dobos and Abonyi, 2012)). Some works are based on the direct use of SPCA, or a hybrid of SPCA with other metrics, in problems involving time series segmentation and classification through snapshot data ((Singhal and Seborg, 2002); (Harrou *et al.*, 2015)). In these cases the patterns are previously chosen by the user or determined based on expert knowledge. Other approaches comprise the application of a clustering algorithm such as the classical FCM algorithm ((Coppi, D'urso and Giordani, 2010); (D'urso, 2004)) together with the use of SPCA with or without other metrics ((Fontes and Budman, 2017); (Fontes and Pereira, 2016); (Izakian, Pedrycz and Jamal, 2015)). Model-based approaches ((Yang and Jianmin Jiang, 2018)) are also used in the clustering of temporal data. However, the time series clustering approaches (raw-data-based, feature-based and model-based, (Liao, 2005), (Aghabozorgi, Shirkhorshid and Wah, 2015)) do not ensure that the recognized patterns are consistent with the dynamic behavior of the process. Depending on the quality of information in the data, this can lead to obtaining unfeasible or non-achievable patterns while obtaining good classification and clustering results.

Constrained clustering (also known as intelligent clustering or semi-supervised clustering) allows previous knowledge to be added by integrating hard or soft constraints into the clustering problem, which can be classified as cluster-level, feature-level or instance-level

constraints (Wagstaff, 2002, Dao et al., 2017, (Lampert *et al.*, 2018)). The first sets conditions for the cluster itself such as the size or the diameter. The second allows some objects to be placed in a given cluster according to their intrinsic features and the third (the most frequently used) specifies that two objects (instances or items) must be or cannot be placed in the same cluster (*must-link* and *cannot-link* constraints respectively). The concepts related to the constraint clustering have already been widely established ((Kiri Lou Wagstaff, 2002), (Law, Topchy and Jain, 2004), (Charikar, Guruswami and Wirth, 2005)). Dao, Duong and Christel Vrain (2017) propose a Constrained Programming framework based on a bi-criterion optimization subject to hard user-constraints. The authors discuss the effects of different criteria, usually conflicting, and the use of Paretto optimal solutions. Others present strategies to include knowledge in specific appliactions such as portfolio optimization, index tracking and direct marketing ((Seret, Verbraken and Baesens, 2014), (Wu, Kwon and Costa, 2017)). Oliveira, Chaves and Lorena (2017) propose two heuristic methods to solve the constrained clustering problem comprising both must-link and cannot-link constraints. The *k*-Means is still the most explored method in these studies ((K. Wagstaff *et al.*, 2001); (Oliveira, Chaves and Lorena, 2017), (Diez-Olivan *et al.*, 2017), (Dao, Duong and Christel Vrain, 2017)).

The application of constrained clustering in time series is still incipient (Lampert *et al.*, 2018), both in whole and subsequence approaches (Aghabozorgi, Shirkhorshid and Wah, 2015), and, as far we know, there are no works related to constrained clustering involving multivariate time series.

Data reconciliation is a well-known and widely adopted technique that relies on minimizing measurement errors in the data by imposing physical constraints associated with the production system (usually mass and energy balances) ((Li *et al.*, 2018), (Syed *et al.*, 2016)). The dynamic behavior of time series collected directly from the process database may present inconsistencies not only due to measurement errors but also to the effect of unmeasured or unknown disturbances intrinsic to the process itself. In this case, considering that the clustering of time series involves an optimization problem to be solved by a classical search method, even the initial guess can effect the quality of the recognized patterns.

This paper presents a new approach which comprises the reconciliation of patterns in the clustering of time series. “Pattern Reconciliation” means recognizing different clusters and patterns in a set of time series extracted from the database of a given production process and at the same time making sure that the recognized patterns are consistent with the dynamic

behavior of this process. The dynamic behavior is represented by a model (hybrid, empirical or phenomenological). The approach should be able to recognize reconciled patterns without worsening the quality of clustering and classification of the objects (time series). Considering the uncertainties involved in defining the boundaries of the clusters, the proposed strategy is based on Fuzzy C-Means (FCM) and directly considers the process dynamics as a soft constraint in order to ensure the feasibility of the recognized patterns. The problem comprises a bi-criterion optimization subject to the hard constraints associated with the process variables.

The contribution of this work can be summarized in the following items:

- For the first time, it is demonstrated that the optimization-based clustering involving multivariate time series, together with a classical similarity metric, can provide patterns distant from the dominant dynamics of the process or even not achievable.
- The specific case of pattern reconciliation involving Univariate Time Series (UTS) is presented and solved analytically. The results obtained are discussed and validated through tailored examples (simple examples).
- Two generic approaches of pattern reconciliation problems involving Multivariate Time Series (MTS) are proposed (both consisting of bi-criterion optimization) according to the nature of the process variables (input or output) considered in each object (MTS). For each type of problem, a generic optimization model with constraints and its resolution strategy (simultaneous or sequential) are proposed.

This paper is structured as follows. Section 2 presents a generic definition for the problem of pattern reconciliation with UTS, its solution based on the first order optimally conditions and a simple example for illustration. Section 3 presents basic definitions about MTS and PCA similarity factor. Two case studies are presented and discussed in Sections 4 and 5. The first case involves the Fault (abnormal operation) Diagnosis in a process consisting of a nonisothermal Continuous Stirred Tank Reactor (CSTR), a well-known benchmark system used for the assessment of FDD techniques. The second case study comprises a real industrial scenario which involves the recognition of starting patterns in a gas turbine for Fault Detection purposes.

2. The Pattern Reconciliation Problem – Univariate Time Series (UTS)

2.1 Fuzzy C-Means Method and optimization-based clustering

Clustering can be categorized as unsupervised learning in which a set of non-labeled objects (data or instances) are divided into homegenous and well separated clusters (subsets)

according to some pre-defined measures of similarity or dissimilarity ((Dao, Duong and Christel Vrain, 2017); (Mitsa, 2010); (Hoppner *et al.*, 1999)). The Fuzzy C-means Method (FCM) is a non-hierarchical method belonging to the C-Means families of batch clustering models ((Bezdek *et al.*, 2005)). Classical FCM is based on an optimization problem Eqs. (1) and (2) whose decision variables are the centers $(v_i, i = 1, \dots, c, \text{prototypes or patterns})$ of each of the c clusters c≥2($c \geq 2$) and the membership degree of each object to each cluster. FCM is suitable for clustering objects represented by vectors in the space $\Re^m$, such as Univariate Time Series (m is the dimensionality) $(x_k, v_i \in \Re^m)$.

$$\underbrace{min}_{U,V} J_\varepsilon(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^{\varepsilon} \cdot \|x_k - v_i\|^2) \quad (1)$$

where c is the number of clusters, n is the number of objects, u_{ik} is the membership degree of the *kth* object to the *ith* cluster, U is the partition matrix ($c \times n$ matrix). The parameter ε $(\varepsilon > 1)$ is the fuzzification coefficient (in this work, $\varepsilon = 2$). V is the set of prototype vectors $\{v_1, \dots, v_c\}$. Two additional constraints related to the membership degrees must be considered:

$$\begin{cases} u_{ik} \in [0,1] \;\; and \;\; \sum_{k=1}^{n} u_{ik} > 0 \;\; \forall i \\ \sum_{i=1}^{c} u_{ik} = 1 \;\; \forall k \text{ (probabilistic approach)} \end{cases} \quad (2)$$

Considering the use of the Euclidean distance as a similarity metric, the application of the first order optimally conditions (necessary conditions) for the problem defined by Eqs. (1) and (2) leads to the following analytical solution:

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^{\varepsilon} \cdot x_k}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} \quad i = 1, \dots, c \quad (3)$$

$$u_{ik} = \frac{\left(\frac{1}{\|x_k - v_i\|^2}\right)^{\frac{1}{\varepsilon-1}}}{\sum_{j=1}^{c}\left(\frac{1}{\|x_k - v_j\|^2}\right)^{\frac{1}{\varepsilon-1}}} \qquad k = 1, \dots, n \quad i = 1, \dots, c \quad (4)$$

2.2 Bi-criterion approach - Fuzzy C-Means and pattern reconciliation

Consider a typical clustering and pattern recognition problem involving Univariate Time Series (UTS). Each object is a time series of length m and a sample with n time series extracted from the historical database of a given process is available. Each object is a vector in $\Re^m$ $(x_k \in \Re^m, k = 1, \dots, n)$. On the other hand, each time series represents the dynamic response, for a given period of time, of a specific process variable (y, process output) due to

the disturbances or changes in other process variables (input variables). *y* is previously defined as able to provide patterns of behavior of the process itself.

Consider that g input variables $(w_l, l = 1, \dots, g)$ have an effect on the output *y* and these effects can be predicted (described) through a dynamic model (empirical, phenomenological or hybrid) of the process. Suppose that this model is known and represented by:

$$\hat{y}(t) = f\left(p, w_1(t), \dots, w_g(t)\right) \quad (5)$$

where $\hat{y}(t)$ is the output predicted by the model, *t* is the continuous time, *p* represents the model parameters (constants as the model is not time-varying) and *f* is the model itself. Output measurements $\left(\hat{y}(t)\right)$ based on the same sampling period adopted in the data collection can be stored in the vector $\hat{y}_c$ $(\hat{y}_c \in \Re^m)$.

Considering *c* clusters, each of the prototype vectors (patterns) $\{v_1, \dots, v_c\}$ is also a dynamic response of the output variable (*y*) and should be feasible, which implies that each pattern should be as close as possible to the behavior predicted by the model and at the same time resulting from feasible changes in the input variables. Thus, the following problem involving bi-criterion constrained clustering is proposed:

$$\underbrace{min}_{U,V^r,W} \; H_\varepsilon(U, V^r, W) = \alpha \cdot \{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^{\varepsilon} \cdot \|x_k - v_i^r\|^2)\} + (1-\alpha) \cdot \left\{\sum_{i=1}^{c}\left\|v_i^r - \hat{y}_{c_i}\right\|^2\right\} \quad (6)$$

Subject to

$$\hat{y}_{c_i} = \left[\hat{y}_{c_i}(1), \dots, \hat{y}_{c_i}(m)\right]^T \quad (7)$$

$$\hat{y}_{c_i}(k_t) = f_d\left(w_{i1}(k_t - dt_1), \dots, w_{i1}(k_t - nw_1), w_{i2}(k_t - dt_2), \dots, w_{i2}(k_t - nw_2), \dots, w_{ig}\left(k_t - dt_g\right), \dots, w_{ig}\left(k_t - nw_g\right)\right) k_t = 1, \dots, m \;\; and \;\; i = 1, \dots, c \quad (8)$$

$$u_{ik} \in [0,1] \;\; and \;\; \sum_{k=1}^{n} u_{ik} > 0 \;\; \forall i \quad (9)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \;\; \forall k \quad (10)$$

$$L_l \le w_{il}(k_t) \le S_l, \quad l = 1, \dots, g \;\; and \;\; i = 1, \dots, c \quad (11)$$

$$\Delta L_l \le \Delta w_{il}(k_t) \le \Delta S_l, \;\; where \;\; \Delta w_{il}(k_t) = w_{il}(k_t) - w_{il}(k_t - 1)\,, \quad l = 1, \dots, g \;\; and \;\; i = 1, \dots, c \quad (12)$$

f_d is the model (Eq. 5) in the discrete form, $dt_l \;\; and \; nw_l$ $(l = 1, \dots, g)$ are the dead time and number of past values, respectively, of each input in the discrete model. k_t is the time instant, $L_l \; and \; S_l$ are the lower and upper limits of each input and $\Delta L_l \;\; and \;\; \Delta S_l$ are the minimum and maximum variations of each input at each time instant. Eqs. (11) and (12) represent the

physical constraints (hard constraints) of the process. $w_{il}(k_t)$ is the value of the input variable l at time instant k_t, associated with the cluster i $(i = 1, \dots, c)$. α $(\alpha \in [0,1])$ is a tuning parameter (trade-off parameter). W is the set of input profiles associated with each cluster $\{w_{il} \in \Re^m, i = 1, \dots, c \ and \ l = 1, \dots, g\}$ and V^r is the set of reconciled patterns $\{v_i^r, i = 1, \dots, c\}$. Each pattern is related to the specific input profiles considering the same model (Eqs. 5 or 7) and the same physical constraints (Eqs. 11 and 12). The dynamic profiles of each input associated with each pattern are also decision variables and each $\hat{y}_{c_i}$ is the dynamic response of the process closest to the reconciled pattern (v_i^r).

As in the previous section, the application of the first order optimal conditions (necessary conditions) for the problem defined by Eqs. 6-12 (see Appendix) leads to the following analytical solution for the patterns (cluster centers) (also considering the Euclidean distance as similarity metric):

$$v_i^r = \frac{\sum_{k=1}^{n}(u_{ik})^{\varepsilon} \cdot x_k}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}+1} + \frac{\hat{y}_{c_i}}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} \ , \quad i = 1, \dots, c (13)$$

In general, the following approximation can be applied:

$$v_i^r \cong v_i + \frac{\hat{y}_{c_i}}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} \ , \quad i = 1, \dots, c$$

$$\text{or} v_i^r - v_i \cong d_i = \frac{\hat{y}_{c_i}}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} (14)$$

$$\text{where } v_i = \frac{\sum_{k=1}^{n}(u_{ik})^{\varepsilon} \cdot x_k}{\sum_{k=1}^{n}(u_{ik})^{\varepsilon}} \quad i = 1, \dots, c (\text{Eq. } 3)$$

Eq. (14) shows that each reconciled pattern can be considered a deviation (d_i) from the respective unreconciled pattern (v_i) (classical problem, Eqs. 1-2). On the other hand, the sum $\sum_{k=1}^{n}(u_{ik})^{\varepsilon}$ (Eq. 14) considers the membership degree of all n objects (the whole sample) to the cluster i. Assuming the use of normalized values for the model output and time series (data), the sum $\sum_{k=1}^{n}(u_{ik})^{\varepsilon}$ would be sufficiently large such that the deviation d_i would tend to zero in most applications involving Univariate Time Series (UTS). This is valid even in cases where cluster i is not well defined as can be shown in the next section. Therefore, it is not expected that patterns reconciled through equations 6-12 are significantly different from the patterns recognized by the application of the classic FCM algorithm (Eqs. 3-4) when dealing with UTS.

Regarding the membership degrees, the application of first order conditions leads to the same solution (Eq. 4) for both classical/unreconciled (Eqs. 1-2) and bi-criterion problems (Eqs. 6-12).

2.2.1 A simple example

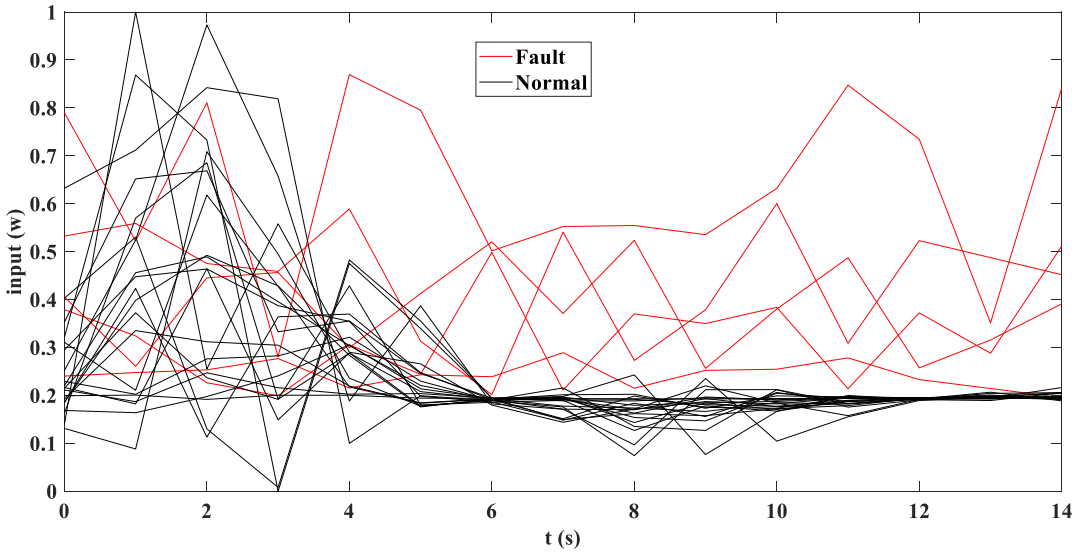
A simple numerical example was tailored to illustrate the conclusions derived from the analysis of Eq. (14).

Consider a hypothetical Single Input Single Output (SISO) process (y *and* w are the output and input variables respectively) whose the dynamic behavior is represented by the following linear AutoRegressive with eXogeneous input (ARX) model:

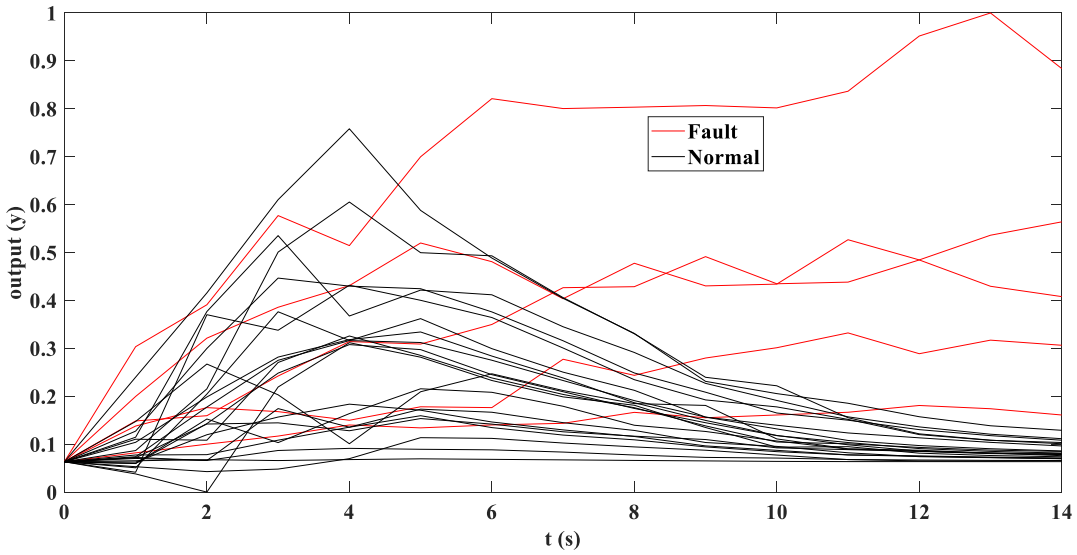
$$y(k_t) + 0.807 \cdot y(k_t - 1) = 0.676 \cdot w(k_t - 1) + 0.005 \cdot w(k_t - 2) + e(k_t) \quad (15)$$

$e(k_t)$ *is white noise disturbance* and $y(k_t - 1), w(k_t - 1), w(k_t - 2)$ are delayed output and input values (regressors).

A sample with 25 objects (25 time series of the output variable collected in different time periods, all of them with the same window length) is available. Suppose that only 5 objects associated with a specific disturbance in the input (classified as a specific failure) are available, and the others are related to the normal operation of the process. Figure 1 presents the time series associated with the input and the process response (output) obtained through the model (Eq. 15).



(a)

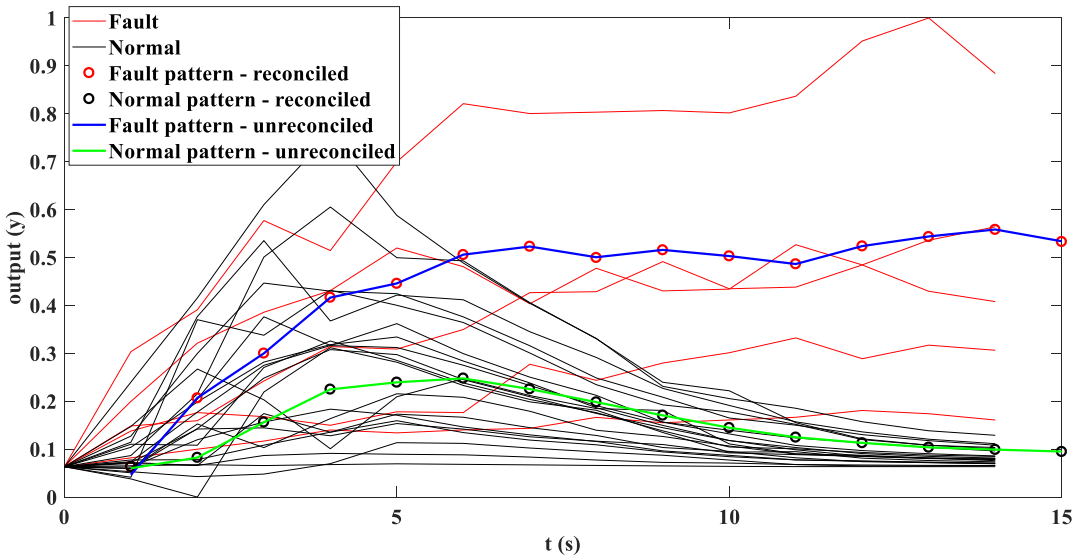


(b)

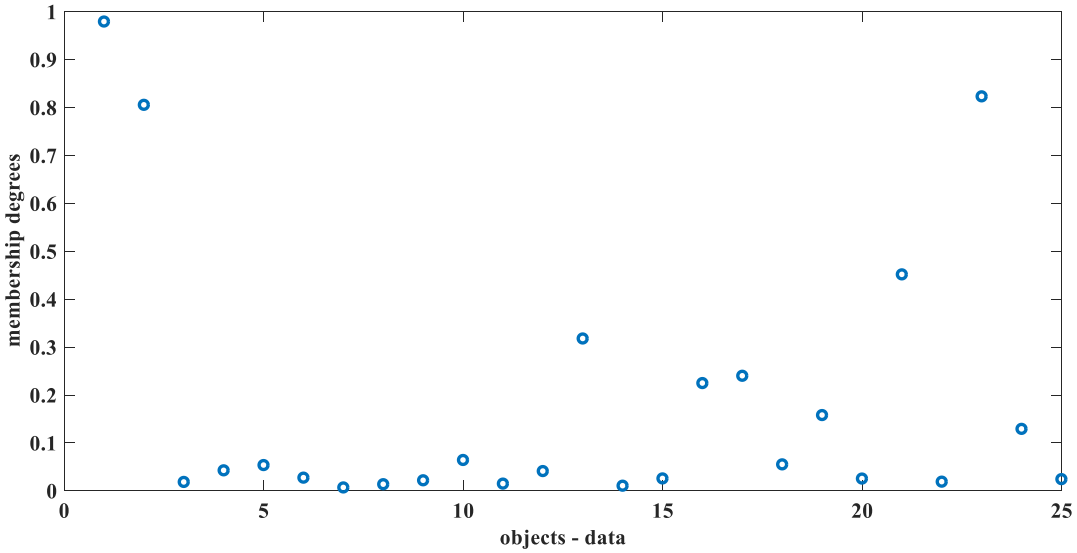
Fig.1. Data sample – time series: a) input b) output

This problem involves unbalanced data sets ((Fontes and Budman, 2017), (Ganganwar, 2012), (CAO *et al.*, 2012) in which the classification problem may be over fitted for the majority class (20 objects, normal operation) leading to a possible difficulty in recognizing the minority cluster (fault).

Figure 2a shows the patterns of each cluster recognized through the classical (Eqs. 1-2 or Eqs. 6-12 with α=1) and bi-criterion approaches (reconciled patterns, Eqs. 6-12, with α=0.5). In both cases, the clustering comprised unsupervised learning in which the label/class of each object (fault or normal) was not informed to the algorithm. The membership degrees of all objects to the fault cluster (minority class) (Figure 2b) shows that only 3 fault objects should be classified as fault objects (2 objects misclassified). In addition, the membership degrees also indicate that the fault cluster is really poorly represented in the sample. On the other hand, the unreconciled and reconciled patterns recognized for each cluster are quite similar. The sum presented in Eq. 14 ($\sum_{k=1}^{n}(u_{ik})^{\varepsilon}$) refers to the membership degrees of all objects and not only objects classified as faults. Therefore, even in problems involving poorly represented classes and with low homogeneity among their objects, it is expected that the reconciliation procedure will not provide a new pattern different from the original one (unreconciled). Only clustering problems involving very small samples could lead to a meaningful difference between reconciled and unreconciled patterns involving UTS.



(a)



(b)

Fig. 2. Output data and patterns reconciled ($\alpha = 0.5$) and unreconciled ($\alpha = 1$); b) membership degrees to the fault cluster (minority class).

3. The Pattern Reconciliation Problem – Multivariate Times Series (MTS)

Multivariate Time Series comprises a collection of two or more time series associated with different process variables and related to the same period of time. The use of MTS is motivated by the need to analyze the problem of classification in a multivariate way and to detect joint features or information hidden in the variables as a whole ((Xun and Zhishu,

2010), (Plant, Wohlschlager and Zherdin, 2009)). Therefore, similarity analysis should not consider dimensionality reduction approaches that may lead to loss of information.

Given a general series of observations over time (time series) associated with a specific process variable (input or output) $\left(z_j(k_t),\ j = 1, \dots, p,\ where\ p \text{ is the number of variables, and } k_t = 1, \dots, m\right)$, an MTS object refers to the case in which $p \geq 2$ and can be represented by the following $m \times p$ matrix:

$$Z_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ip}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ip}(m) \end{bmatrix} \tag{15}$$

Z_i is the object, $z_{ij}(k_t)$ is the measurement of variable j at time instant k_t in the object Z_i $(i = 1, \dots, n)$. Each column contains the time series related to a given variable.

The PCA (Principal Component Analysis) Similarity metric (SPCA) ((Yang and Shahabi, 2004), (Dobos and Abonyi, 2012)) is a well-known measure of similarity between two different MTS with the same number of variables (p) but not necessarily the same length of time window $(\text{number of observations}, m)$. The SPCA index measures the similarity between two MTS through the similarity between the directions of its principal components.

Consider that the number of principal components associated with the objects (matrices) $Z_A\ and\ Z_B$ $(\text{respectively } k_A\ and\ k_B)$ is capable of representing at least 95% of the total variance in each object. The variance related to each component can be computed directly by the eigenvalues of the covariance matrix associated to each MTS. Since the principal components describe different variances in data, a modified version of the original SPCA index $(SPCA_\lambda)$ in which each principal component is weighted by the square root of its corresponding eigenvalue is more appropriate (Singhal and Seborg, 2006):

$$SPCA_\lambda(Z_A, Z_B) = \frac{1}{\sum_{i=1}^{k_0}(\lambda_i^A \cdot \lambda_i^B)} \cdot \sum_{i=1}^{k_0}\sum_{j=1}^{k_0}\left(\lambda_i^A \cdot \lambda_j^B\right) \cdot \left(cos\theta_{ij}\right)^2 \tag{16}$$

k_0 is set as the largest of $k_A\ and\ k_B$, $\lambda^A\ and\ \lambda^B$ are vectors with the eigenvalues of $Z_A^T \cdot Z_A$ and $Z_B^T \cdot Z_B$, respectively. θ_{ij} is the angle between the *ith* principal component of Z_A and the *jth* principal component of Z_B. The PCA performed for each time series is mean centered.

3.1 MTS with output variables

The first type of problem involving MTS comprises the case in which each time series present in the object is associated with the dynamic behavior of an output variable. In this case, a

sequential optimization approach is proposed in order to solve the problem of pattern reconnection.

Each pattern (center of each cluster) is also a MTS and the first step comprises only the clustering itself, according to the following bi-criterion approach:

$$\underbrace{min}_{U,V}\ \Omega_{\varepsilon}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon}\cdot\left(SPCA_c(X_k,V_i)\right)^2\right) + \beta\cdot\sum_{i=1}^{c}\sum_{\substack{j=1\\ j>i}}^{c}\frac{1}{\left(SPCA_c(V_j,V_i)\right)^2} \quad (17)$$

Subject to

$$\begin{cases} u_{ik}\in[0,1] \ \ and \ \ \sum_{k=1}^{n}u_{ik} > 0 \ \ \forall i \\ \sum_{i=1}^{c}u_{ik} = 1 \ \ \forall k \end{cases} \quad (18)$$

where

$$SPCA_c(X_k,V_i) = 1 - SPCA_{\lambda}(X_k,V_i) \quad (19)$$

$SPCA_c$ is the complement of $SPCA_{\lambda}$ (Eq. 16) since Eqs. 17-18 involving a minimization problem and values of $SPCA_{\lambda}$ close to one imply high similarity.

V is the set of patterns, $X_k\ (k=1,\dots,n)$ is an object (MTS), $V_i(i=1,\dots,c)$ is a pattern $\left(X_k\ and\ V_i\in \Re^{m\times p}\right)$. $SPCA_c(X_k,V_i)$ is the distance between an object and the pattern (center) based on the modified SPCA (Eqs. 16 and 19).

The second criterion in Eq. 17 (weighted by the parameter β) is designed to avoid the recognition of very close patterns which would imply the overlapping of clusters. This criterion indirectly provides an alternative way to maximize the minimal split between clusters (Dao, Duong and Christel Vrain, 2017) or can also be considered as a constrained on the mining/clustering model ((Grossi, Romei and Turini, 2017)).

Once the patterns have been recognized through Eqs. 17-18, reconciliation is accomplished through another optimization problem that aims to approximate the dynamic profile of each time series to the features of the process, based on its dynamic model and physical constraints. Considering that each V_i obtained as a solution from the previous step (Eqs. 17-18) can be represented by

$$V_i = \begin{bmatrix} v_{i1}(1) & \cdots & v_{ip}(1) \\ \vdots & \ddots & \vdots \\ v_{i1}(m) & \cdots & v_{ip}(m) \end{bmatrix} = \left\{V_{i1},\ V_{i2},\dots,\ V_{ip}\right\} \quad (20)$$

Where each V_{ij} is the time series related to the variable $j\ (j=1,\dots,p)$ in the pattern $(i=1,\dots,c)\ \left(V_{ij}\in \Re^{m}\right)$.

The second optimization problem is

$$\underbrace{\min}_{W,V^r} \Pi_\varepsilon(V^r, W) = \sum_{i=1}^{c}\sum_{j=1}^{p}\left\|V_{ij}^r - \hat{y}_{c_{ij}}\right\|^2 \quad (21)$$

Subject to

$$\hat{y}_{c_{ij}} = \left[\hat{y}_{c_{ij}}(1), \dots, \hat{y}_{c_{ij}}(m)\right]^T \quad (22)$$

$$\hat{y}_{c_{ij}}(k_t) = f_{dj}\left(w_{i1}(k_t - dt_1), \dots, w_{i1}(k_t - nw_1), w_{i2}(k_t - dt_2), \dots, w_{i2}(k_t - nw_2), \dots, w_{ig}(k_t - dt_g), \dots, w_{ig}(k_t - nw_g)\right); k_t = 1, \dots, m;\ i = 1, \dots, c \ \text{and}\ j = 1, \dots, p \quad (23)$$

$$L_l \le w_{il}(k_t) \le S_l, \quad l = 1, \dots, g \ \text{and}\ i = 1, \dots, c \quad (24)$$

$$\Delta L_l \le \Delta w_{il}(k_t) \le \Delta S_l, \quad where\ \Delta w_{il}(k_t) = w_{il}(k_t) - w_{il}(k_t - 1)\,, \quad l = 1, \dots, g \ \text{and} \ i = 1, \dots, c \quad (25)$$

Eqs. 21-25 present a reconciliation approach for the patterns consisting of only output variables and described by MTS.

f_{dj} is the process model in the discrete form that relates the output j $(j = 1, \dots, p)$ to the process inputs, dt_l and nw_l $(l = 1, \dots, g)$ are the dead time and number of past values, respectively, of each input in the discrete model. k_t is the time instant, L_l, S_l, ΔL_l and ΔS_l are the same as Eqs. 11-12 which also represent physical constraints (hard constraints) associated with the input variables. $w_{il}(k_t)$ is the value of the input variable *l* at time instant k_t, associated with the cluster i $(i = 1, \dots, c)$. *W* is the set of input profiles associated with each cluster $\{w_{il} \in \Re^m, i = 1, \dots, c \ \text{and}\ l = 1, \dots, g\}$ and V^r is the set (MTS) of reconciled patterns $\left(V_{ij}^r \in \Re^m, i = 1, \dots, c,\ j = 1, \dots, p\right)$. $\hat{y}_{c_{ij}}$ $\left(\hat{y}_{c_{ij}} \in \Re^m\right)$ is the dynamic response of the output j closest to its reconciled pattern in cluster i $\left(V_{ij}^r\right)$.

The metric distance used in Eq. 21 is the Euclidean distance in order to reconcile the dynamics of the output to the dominant dynamics of the process (predicted by the process model).

3.2 MTS with input and output variables

The second type of problem involving MTS comprises the case in which each object consists of some time series associated with the inputs and others associated with the outputs. The

reconciliation correlates each output in each pattern with the inputs (in the same pattern) according to the process model, without worsening the quality of clustering and classification. Consider a process with a total of g input variables, of which g_1 inputs $(g_1 < g)$ are part of each object and g_2 are the remaining inputs $(g_1 + g_2 = g)$. The first ones are called **Prototype Inputs** (PI) and the others are called **Non-Prototype Inputs** (NPI). Therefore, the set of input variables is divided into two subsets, namely, w_j^{PI}, $(j = 1, \dots, g_1)$ and w_j^{NPI}, $(j = 1, \dots, g_2)$.

Each recognized pattern (and each object) $V_i\,(i = 1, \dots, c)$ is composed of time series referring to p outputs and g_1 Prototype Inputs, comprising a matrix with $p + g_1$ columns. $V_{ij}^{y} \in \Re^{m}$ $(i = 1, \dots, c$ and $j = 1, \dots, p)$ refers to the time series of the jth output in the pattern associated with the cluster i and w_{ij}^{PI} $(i = 1, \dots, c$ and $j = 1, \dots, g_1)$ refers to the time series of the *jth* PI in the pattern associated with the cluster i.

$$V_i = \left\{ \underbrace{V_{i1}^{y}, \dots, V_{ip}^{y}}_{p\ outputs}\ \underbrace{w_{i1}^{PI}, \dots, w_{ig_1}^{PI}}_{\substack{g_1\ prototype \\ inputs}} \right\} \tag{26}$$

The reconciliation strategy is carried out in this case through a simultaneous approach. A bi-criterion constrained clustering (Eq. 27) comprises an objective function whose the first criterion is related to the fuzzy clustering itself and the second imposes a relationship between the time series of each output and prototype inputs belonging to the same pattern.

$$\underbrace{min}_{U,V,W}\ \rho_{\varepsilon}(U,V,W) = \alpha \cdot \left(\underbrace{\sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon}\cdot\left(SPCA_c(X_k,V_i)\right)^2\right) + \beta\cdot\sum_{i=1}^{c}\sum_{\substack{j=1 \\ j>i}}^{c}\frac{1}{\left(SPCA_c(V_j,V_i)\right)^2}}_{clustering} \right) +$$

$$\underbrace{(1-\alpha)\cdot\sum_{i=1}^{c}\sum_{j=1}^{p}\left\|V_{i1}^{y} - \hat{y}_{c_{ij}}\right\|^2}_{pattern\ reconciliation} \tag{27}$$

Subject to

$$\begin{cases} u_{ik} \in [0,1] \ \ and \ \ \sum_{k=1}^{n} u_{ik} > 0 \ \ \forall i \\ \sum_{i=1}^{c} u_{ik} = 1 \ \ \forall k \end{cases} \tag{28}$$

$$\hat{y}_{c_{ij}} = \left[\hat{y}_{c_{ij}}(1), \dots, \hat{y}_{c_{ij}}(m)\right]^T \tag{29}$$

$$\hat{y}_{c_{ij}}(k_t) = f_{dj}\left(w_{i1}^{PI}(k_t - dt_1^{PI}), \dots, w_{i1}^{PI}(k_t - nw_1^{PI}), , \dots, w_{ig_1}\left(k_t - dt_{g_1}^{PI}\right), \dots, w_{ig_1}\left(k_t - nw_{g_1}^{PI}\right), w_{i1}^{NPI}(k_t - dt_1^{NPI}), \dots, w_{i1}^{NPI}(k_t - nw_1^{NPI}), , \dots, w_{ig_2}\left(k_t - dt_{g_2}^{NPI}\right), \dots, w_{ig_2}\left(k_t - nw_{g_2}^{NPI}\right)\right); k_t = 1, \dots, m;\ \ i = 1, \dots, c \ \ and \ \ j = 1, \dots, p \quad (30)$$

$$L_l \leq w_{il}(k_t) \leq S_l, \quad l = 1, \dots, g \ \ and \ \ i = 1, \dots, c \quad (31)$$

$$\Delta L_l \leq \Delta w_{il}(k_t) \leq \Delta S_l, \ \ where \ \ \Delta w_{il}(k_t) = w_{il}(k_t) - w_{il}(k_t - 1)\ ,$$
$$l = 1, \dots, g \ \ and \ \ \ \ i = 1, \dots, c \quad (32)$$

The physical constraints (Eqs. 31 and 32) apply for all inputs, regardless of their feature (PI or FI). Likewise, each model must consider the effect of all inputs (Eq. 30). $dt_j^{PI}\ and\ nw_j^{PI}$ $\left(dt_j^{NPI}\ and\ nw_j^{NPI}\right)$ are the dead time and number of past values of *jth* PI (NPI) in the process model.

As in previous problems, *W* is the set of input profiles associated with each cluster $\{w_{il} \in \Re^m, i = 1, \dots, c \ \ and \ \ l = 1, \dots, g\}$. The dynamic behavior of non-prototype inputs can be pre-defined or not in the optimization (Eqs. 27-32) depending on the information regarding the real problem. If a given NPI is pre-defined (pre-set), this will not be a decision variable.

$\|\ \ \|$ in Eq. 27 also refers to Euclidean distance.

4. Case Studies and Results

4.1 Simulation case study – Continuous Stirred Tank Reactor

The dynamics of a non-isothermal Continuous Stirred Tank Reactor (CSTR) with cooling jacket and a variable liquid level represents a well-known benchmark process used to compare and/or analyze FDD approaches ((Singhal and Seborg, 2002), (Vaidyanathan and Venkatasubramanian, 1992)). Figure 3 presents the CSTR and feedback control system. There are two control loops, namely, the level control whose the manipulated input is the flow rate of the outlet stream $(Q,\ L/min)$ and the temperature control which consists of a cascade control whose the manipulated variable is the coolant flow rate $(Q_c, L/min)$. Assuming a classical first-order irreversible reaction $(A \rightarrow B,$ $A\ is\ the\ reactant\ and\ B\ is\ the\ \ product)$, a complete phenomenological model based on the mass, energy and component balances is presented in (Singhal and Seborg, 2002). These comprise a system with four ordinary differential equations, four dependent (state) variables

(concentration of species A in the reactor, $C_A(t)\ (mol/L)$, reactor temperature, $T(t)\ (K)$, temperature of the coolant in the cooling jacket, $T_c(t)\ (K)$, and liquid level, $h(t)\ (dm)$).

This case study comprised the diagnosis of two types of failures, both related to disturbances in the flow rate of the feed stream $(Q_F, L/min)$. 30 fault objects are related to step changes with different amplitudes (**Fault 1**) and 30 objects are related to damped and sustained oscillations in the feed stream (**Fault 2**) (oscillations were generated considering different amplitudes and frequencies). In all the simulations carried out to obtain the data (objects), the steady state associated with the nominal operating conditions was considered as the starting point (initial steady state) (Singhal and Seborg, 2002) $(Q_F = 100\,L/min\,,\ T = 402.35\ K\ and\ C_A = 0.037\ mol/L)$.

Each object comprises two time series related to the same period of time and associated with two state variables, namely, reactant concentration in the reactor (C_A) and reactor temperature (T). The problem consists of recognizing different fault patterns (faults 1 and 2) and clustering the objects according to the type of failure, considering an unsupervised learning based on the dynamic behavior of two output variables (C_A, T). Figure 4 presents the whole sample, i.e. time series (time window and sampling period equal to 2 min and 5 s, respectively) associated with the outputs. Each time series was normalized within the range [0;+1] considering the maximum and minimum values of the respective process variable along the entire sample (set of objects). Then, each time series was represented in deviation variable (**dv**) form with respect to its initial value, i.e. the value at the beginning of the time window. The use of a deviation variable allows the generalization of recognized patterns to other operating conditions (other initial steady states). The transformation to deviation variable consists of subtracting all points of the time series from its initial value.

A damped sinusoidal signal was added to the concentration curves related to the oscillatory failure (Fault 2) in order to simulate the presence of an unknown disturbance associated with the process or measuring mechanism.

Figure 4c shows a box-plot analysis (Fault 1 and Fault 2 are referred to by the numbers 1 and 2 respectively). Each box is defined by the mean and variance of the $SPCA_c$ metric considering three cases: i- between Fault 1 objects (1 and 1), ii- between objects from the entire sample (faults 1 and 2) and iii- between Fault 2 objects (2 and 2). The box-plot analysis shows high uniformity between the objects associated with the stepping-type fault (Fault 1) and an apparent ease of visual distinction between stepping-type faults (Fault 1) and

oscillation-type faults (Fault 2), which can be verified through greater dispersion in the distribution of distances between objects from the entire sample (1-2).

This case study aims to compare the quality of the classification obtained by the FCM method (just based on SPCA metric),with and without pattern reconciliation, and to verify the dynamic consistency of the recognized patterns for each cluster (cluster 1, type 1 failure, and cluster 2, type 2 failure).

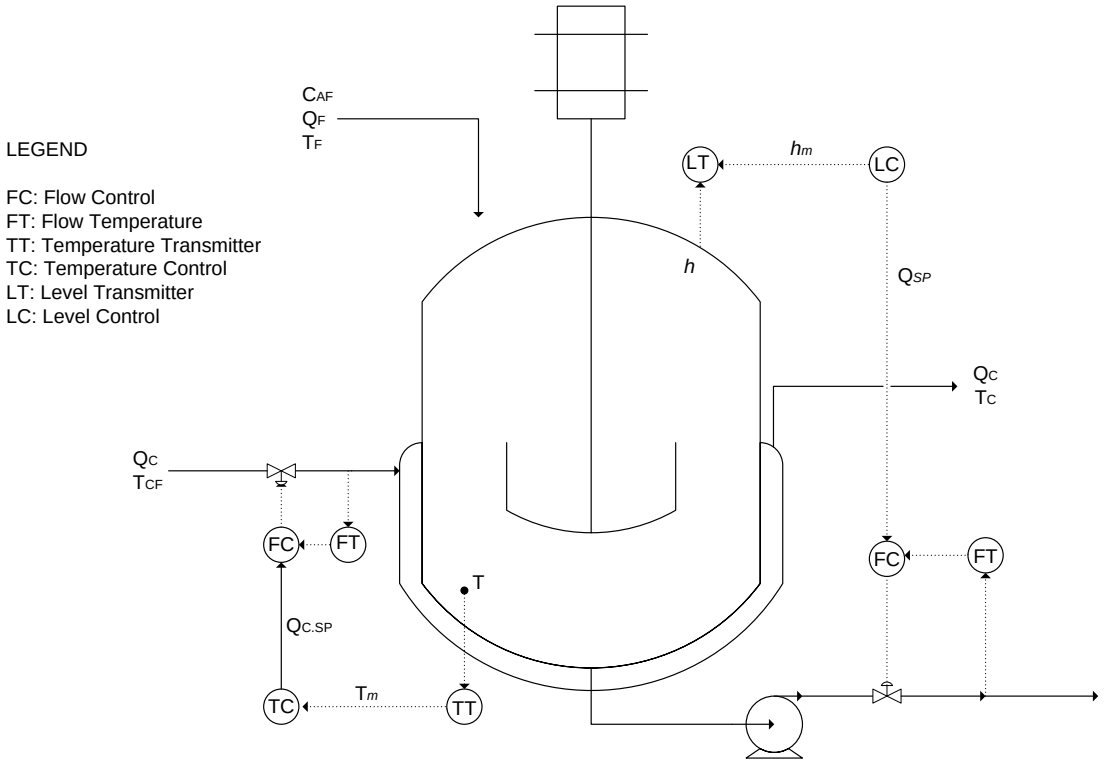
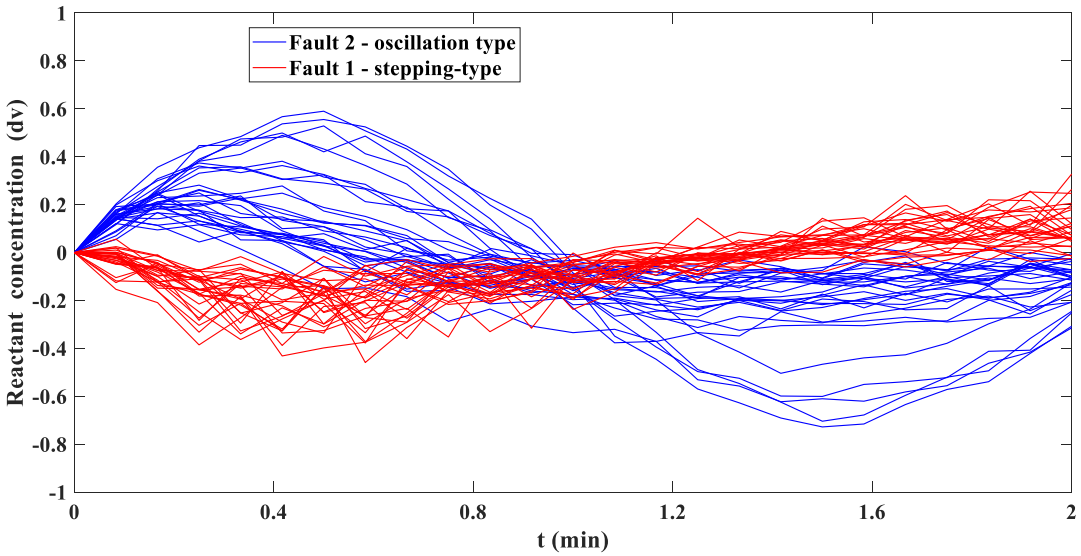
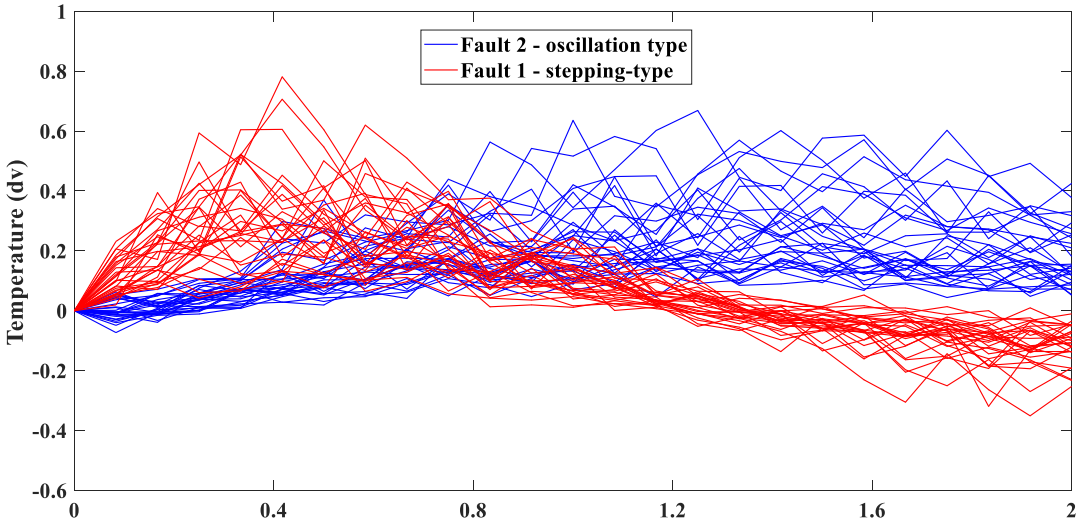


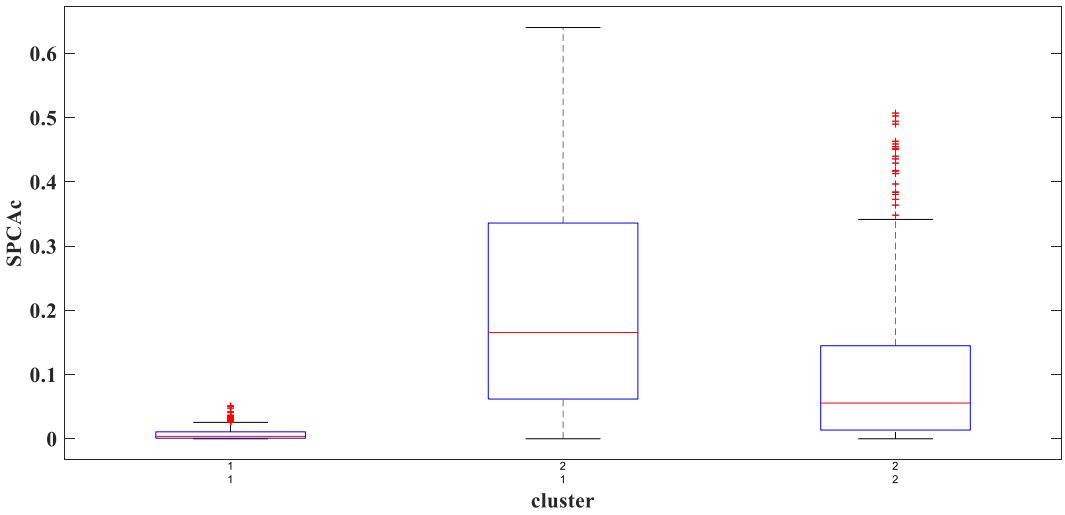
Fig.3. Non-isothermal Continuous Stirred Tank Reactor (CSTR) and control loops.


Fault 2 - oscillation type
Fault 1 - stepping-type
Reactant concentration (dv)
t (min)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
0.4
0.8
1.2
1.6
2

(a)


Fault 2 - oscillation type
Fault 1 - stepping-type
Temperature (dv)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
0.4
0.8
1.2
1.6
2

(b)



(c)

Fig.4. Data sample – Time series: a) reactant concentration (C_A); b) reactor temperature (T) and c) similarities distribution within and between labeled objects (box-plot).

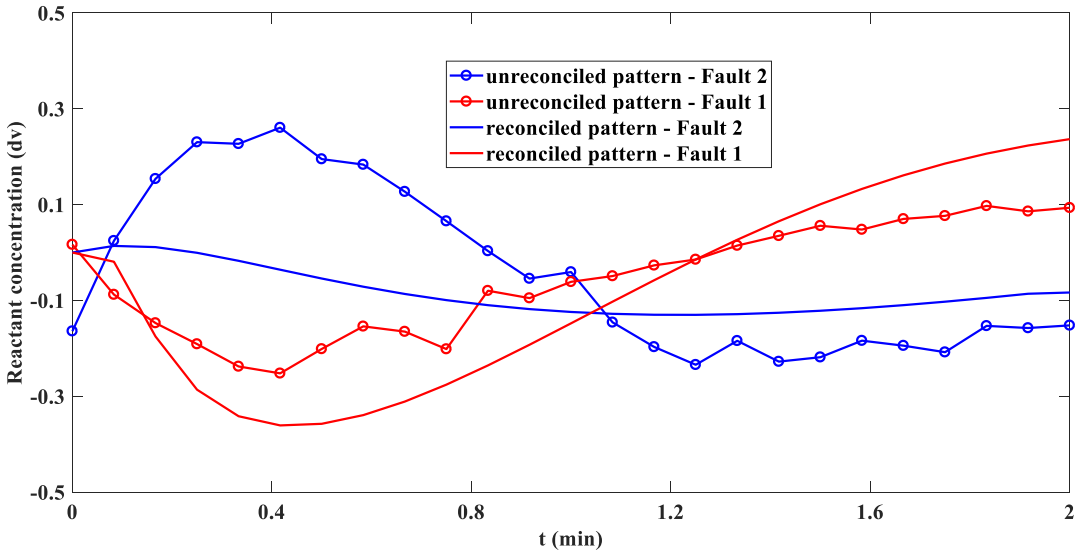
In this case, the clustering and reconciliation involve MTS with only output variables and the procedure consists of solving two optimization problems through a sequential approach (Eqs. 17-18 and 20-25). Table 1 presents the percentage of misclassifications obtained using both approaches (classical FCM without reconciliation, Eqs. 17-18, and sequential approach with pattern reconciliation, Eq.s 17-18 and Eq.s 21-25). Although 3 objects associated with Fault 1 (stepping-type) were misclassified with the reconciled patterns, these same patterns were also able to improve the quality of the classification in relation to the objects of Fault 2 (oscillation-type) which shows that the reconciliation did not affect the quality of the clustering.

Table 1 – Percentage of misclassifications (CSTR, MTS with only output variables)

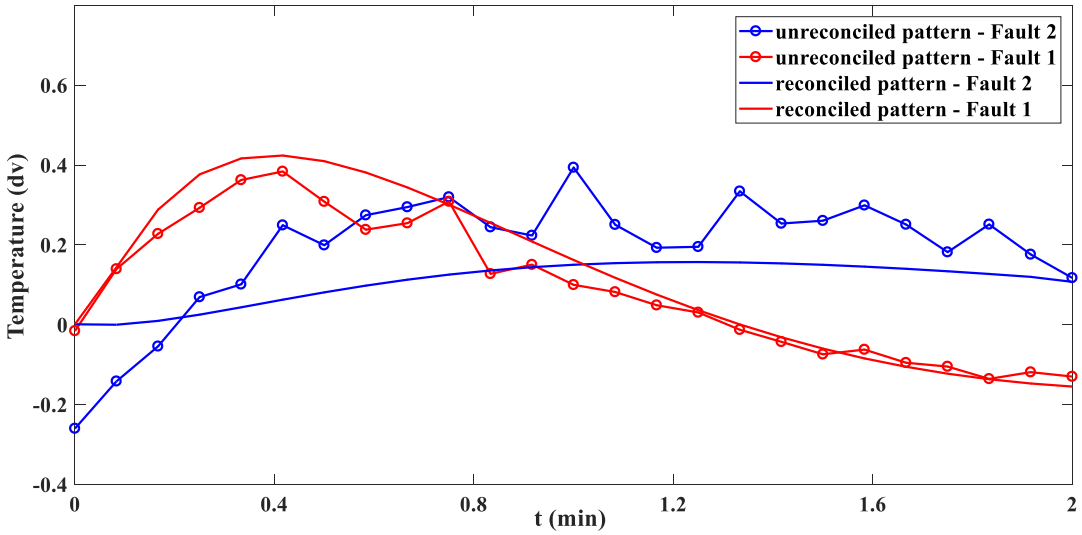
	(unreconciled patterns)	(reconciled patterns)
stepping-type fault (Fault 1)	0%	10% (3 objects)
oscillation-type fault (Fault 2)	17% (5 objects)	3% (1 object)

Figure 5 presents the pattern of each cluster (Fault 1 and Fault 2). The patterns associated with both outputs $\left(C_A(t), T(t)\right)$ are dynamically quite similar especially for the cluster related to Fault 1 (stepping-type), which can be attributed to the greater homogeneity among the time series belonging to this type of failure (Figure 4c). Comparing the patterns of the reactant

concentration recognized for the oscillatory disturbance (Fault 2), a dynamic inconsistency in the behavior of the unreconciled pattern (Figure 5a) can be observed. The initial increase in the reactor temperature (Figure 5b) (caused by the disturbance) should be associated with an initial decrease in the reactant concentration (increase in reactant consumption), which was not found in the unreconciled pattern (the increase in temperature raises the reaction rate and the consumption of the reactant). In addition, the subsequent reduction in the reactant concentration (from about 0.5 min) is not expected since there was no significant increase in temperature pattern. Both reconciled profiles (temperature and reactant concentration) associated with Fault 2 (oscillation-type) present joint behavior consistent with the process showing that the reconciliation was able to eliminate the dynamic effect of the damped sinusoidal noise inserted in the time series of the reactant concentration (Figure 4a). These results show that just the SPCA metric and the classical FCM method were not able to ensure the coherence of the recognized pattern with the process itself, even when there is a small heterogeneity among the objects of the same class (Figure 4c). On the other hand, the reconciled patterns are feasible and also capable of preserving the quality of the classification. The reconciliation approach allows the recognized pattern to be associated with the dominant dynamics of the process not only by inserting the process model into the problem but also by imposing physical constraints related to the input variables (Eqs. 24-25, 31-32). These constraints are also part of the process model and help to avoid the transfer of undesirable noise from the original sample to the pattern.



(a)



(b)

Fig. 5. Reconciled (sequential reconciliation approach) and unreconciled patterns: a) reactant concentration (C_A) and b) reactor temperature (T). Fault 1 (stepping-type) and Fault 2 (oscillation-type).

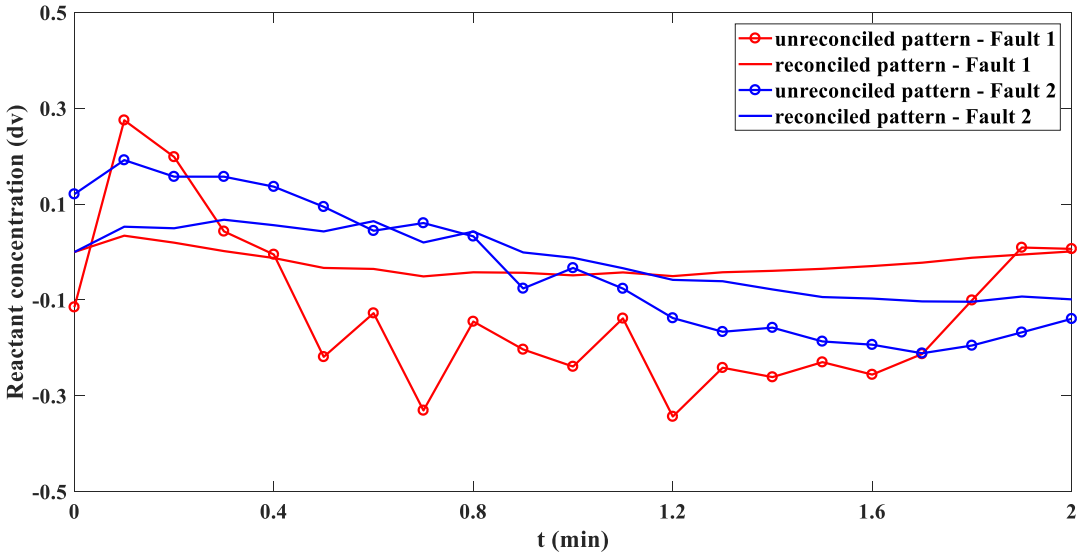
In a second test, each MTS involved the same outputs (C_A, T) as previously and the flow rate of the feed stream (Q_F) (input). The simultaneous reconciliation approach was applied (Eqs. 27-32). Figure 6 presents the unreconciled (Eqs. 17-18) and reconciled patterns, and the time series of feed flow (Q_F) used in the training data (Figure 6d) (together with the time series of temperature and reactant concentration, Figure 4a-b). A damped sinusoidal signal was added to the feed flow curves associated with the step
failure (Fault 1).

Table 2 shows that both approaches (classical FCM without reconciliation, Eqs. 17-18, and simultaneous reconciliation, Eqs. 27-32) obtained the same classification result with only 3 objects associated with Fault 1 (stepping-type) misclassified. Compared to the previous test, both sets of patterns (reconciled and unreconciled) were able to provide a good clustering result which shows that the inclusion of the feed flow $(input\ Q_F)$ in each object (each MTS) makes it easier to diagnose or recognize the type of failure, which in turn can be attributed to the difference between the dynamic behaviors associated with the two types of disturbances considered (step and oscillation, Figure 6d). Furthermore this shows that patterns with different dynamic profiles are capable of providing the similar classification results and, as in

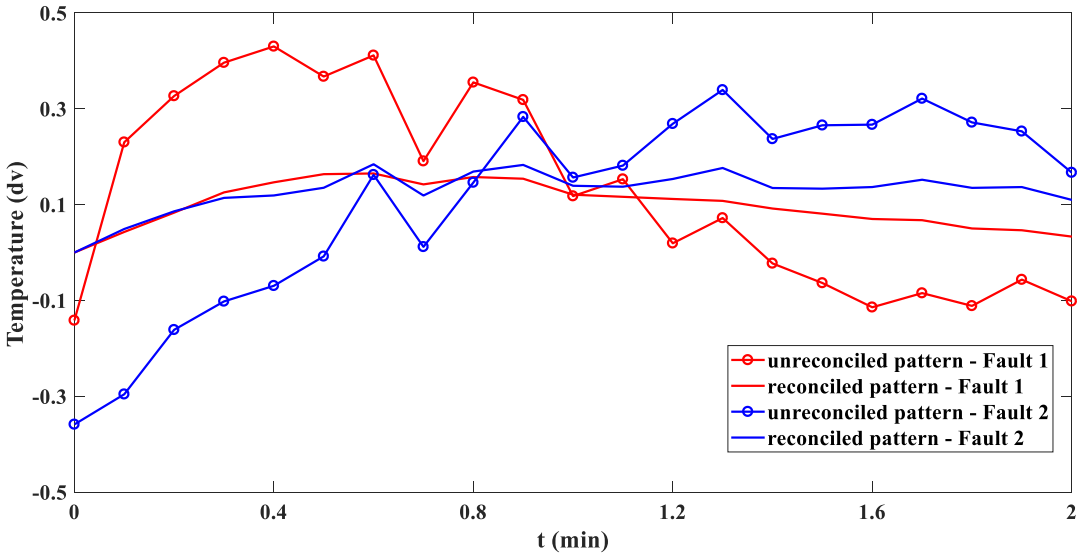
this case, the quality of the clustering should be associated with the feasibility of the recognized patterns. A pattern represented by an MTS is feasible if the time series associated with the input and output variables have interdependent dynamic behaviors predicted by the process model.

Table 2 – Percentage of misclassifications (CSTR, MTS with input and output variables)

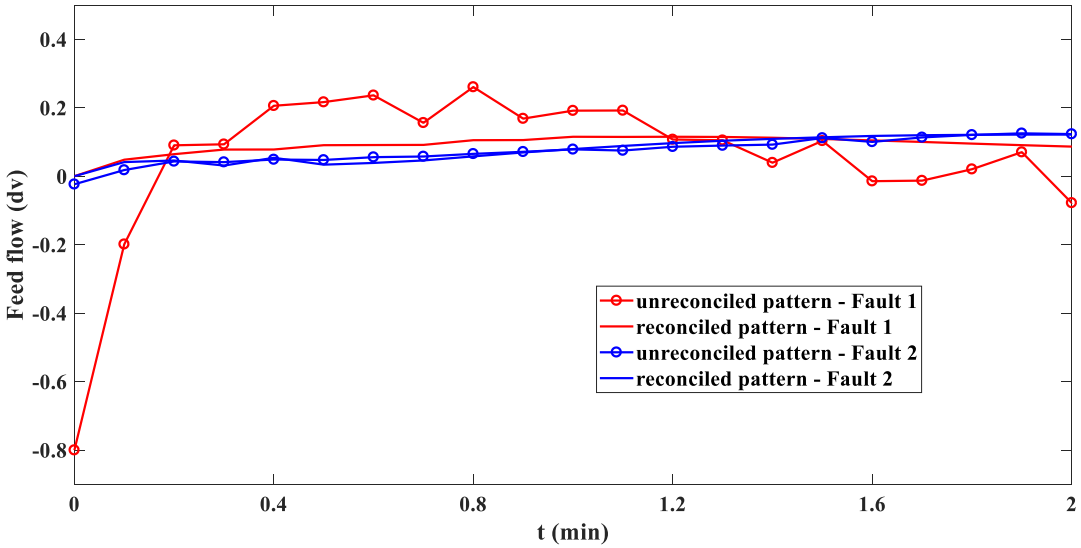
	(unreconciled patterns)	(reconciled patterns)
stepping-type fault (Fault 1)	10% (3 objects)	10% (3 objects)
oscillation-type fault (Fault 2)	0%	0%



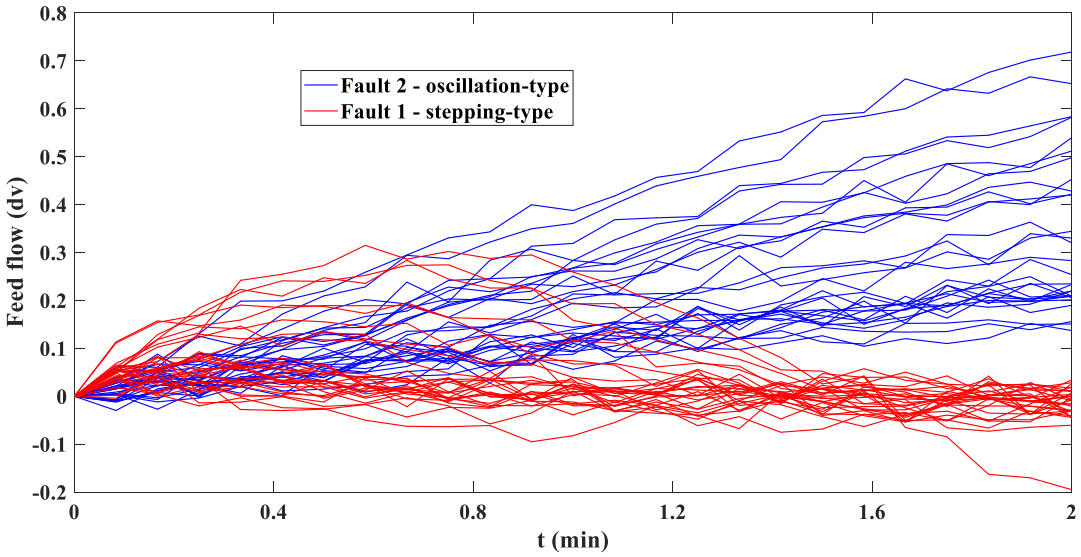
(a)


0.5
0.3
0.1
-0.1
-0.3
-0.5
Temperature (dv)
0
0.4
0.8
1.2
1.6
2
t (min)
unreconciled pattern - Fault 1
reconciled pattern - Fault 1
unreconciled pattern - Fault 2
reconciled pattern - Fault 2

(b)


0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Feed flow (dv)
0
0.4
0.8
1.2
1.6
2
t (min)
unreconciled pattern - Fault 1
reconciled pattern - Fault 1
unreconciled pattern - Fault 2
reconciled pattern - Fault 2

(c)



(d)

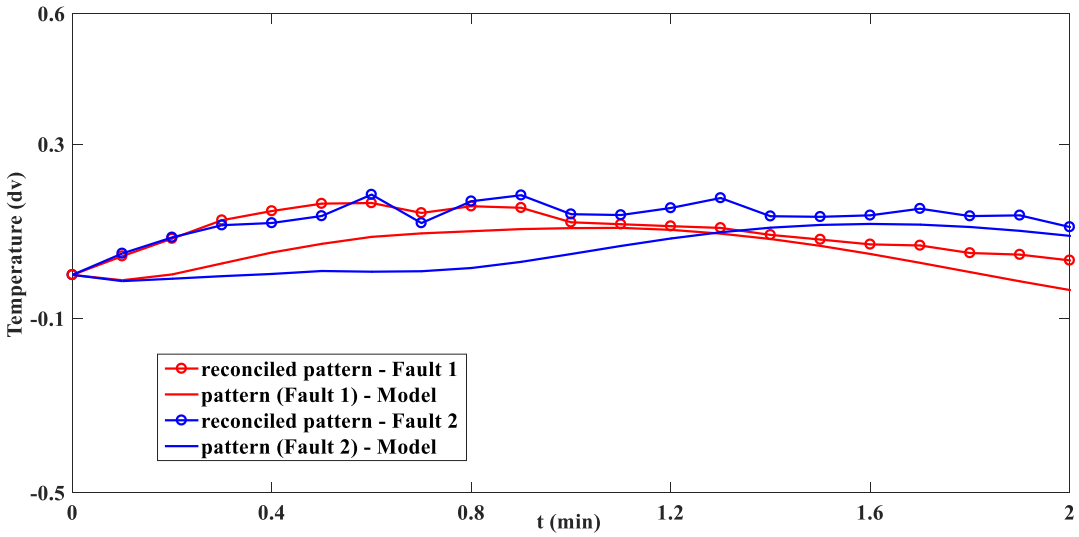
Fig. 6. Reconciled (sequential reconciliation approach) and unreconciled patterns: a) reactant concentration (C_A), b) reactor temperature (T) and c) feed flow (Q_F) d) Time series – feed flow (Q_F).

Because the reconciliation associates the pattern recognized by the FCM method to the dominant dynamics of the process (represented by its model), Figures 6a and 6b show that the patterns obtained in this case tend to be less noisy and capture the deterministic behavior of the process.

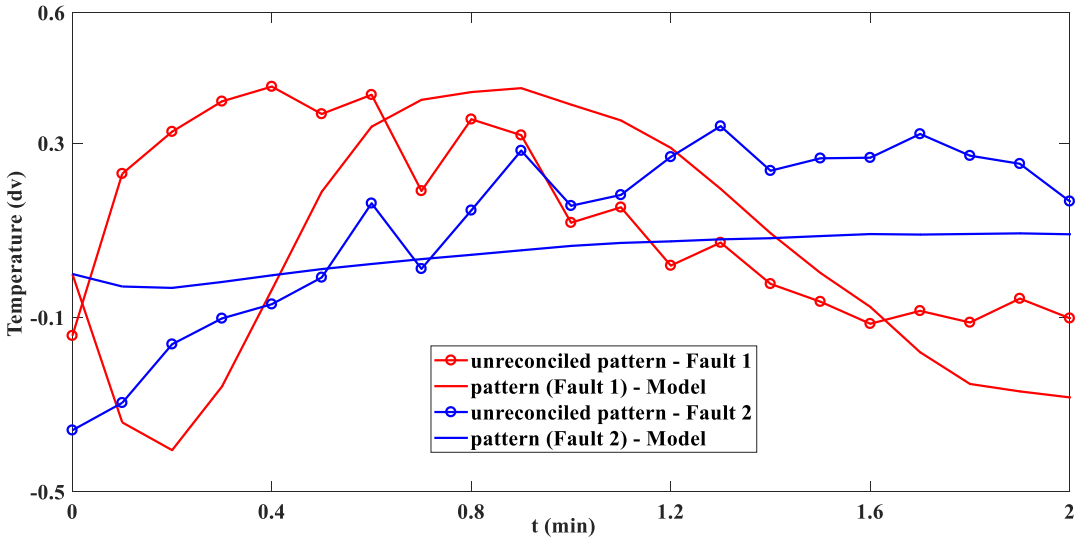
Figure 7 presents the analysis of dynamic feasibility of each pattern (reconciled and unreconciled). Each figure shows the recognized pattern for a temperature (Figures 7a-b) or concentration (Figures 8a-b) and the time series of temperature or concentration predicted by the model based on the respective feed flow curve (Q_F) of each recognized pattern (Figure 6c). In the model simulation, the other input variables were set at their normal operating values (inlet coolant temperature, T_{CF}, concentration of species A in the feed stream, C_{AF}, reactor feed temperature, T_F) (Singhal and Seborg, 2002) or determined automatically by the control loops (coolant flow rate, Q_C, and flow rate of the outlet stream, Q) whose settings remained unchanged.

The similarity between the curves predicted by the model for both outputs (C_A, T) in Fault 2 is expected because the flow curves in each pattern are close enough (Figure 6c). Figure 7 shows that the dynamic behavior of the feed flow (Q_F) is not consistent with those of the

output variables (C_A, T) in the unreconciled patterns. In this case, equal signs of static gains (both positive) show an increase in both outputs at the beginning of the time window (0-0.4 min) (Figures 7b and 8b, Unreconciled patterns, Fault 1). This reveals a dynamic inconsistency between temperature and reactant concentration considering that their respective dynamic profiles are related to the same pattern and to the same dynamic profile of the feed rate (Q_F). Therefore, it can be concluded that the unreconciled pattern associated with failure 1 is not feasible, even though a good classification result (Table 2) was obtained.

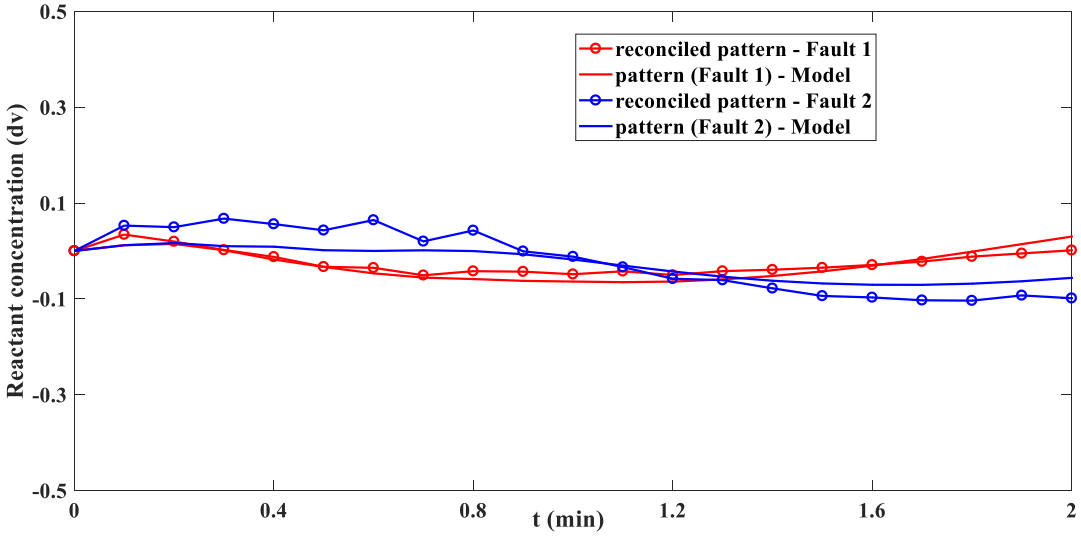


(a)

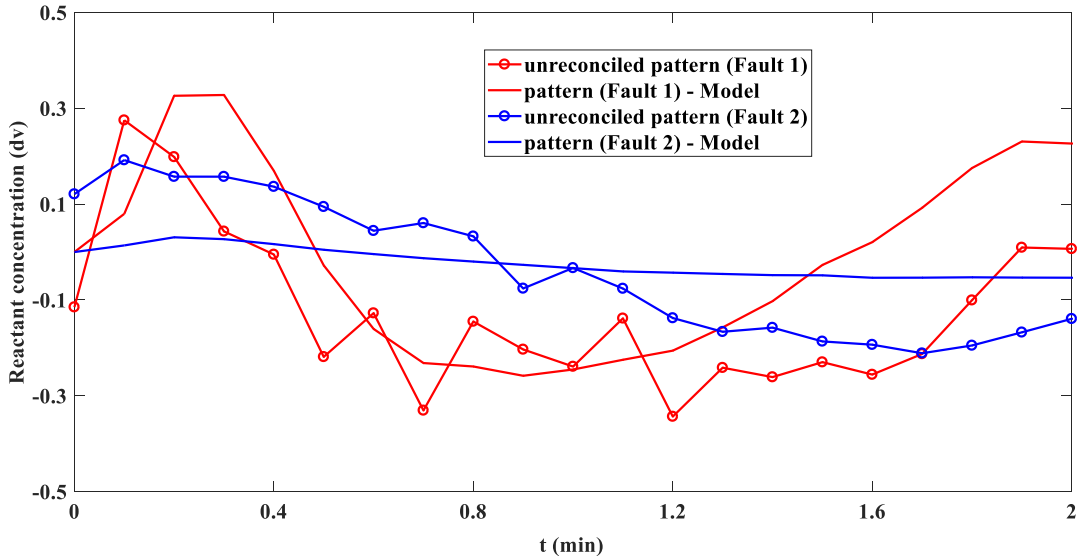


(b)

Fig. 7. Temperature – patterns recognized and predicted by the process model. (a) reconciled and (b) unreconciled. Fault 1 (stepping-type) and Fault 2 (oscillation-type).



(a)



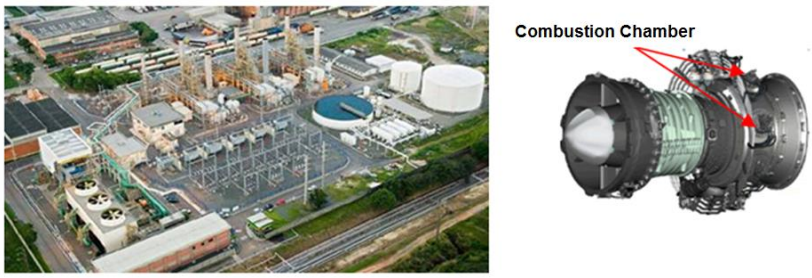
(b)

Fig. 8. Reactant concentration – patterns recognized and predicted by the process model. (a) reconciled and (b) unreconciled. Fault 1 (stepping-type) and Fault 2 (oscillation-type).

4.2 An industrial case study – start-up of a gas turbine

The second case study comprised the application of optimization-based clustering, based on historical data, in a Thermoelectric Power Plant (TPP). It consists of a cogeneration unit that

operates in a combined cycle producing steam and electricity using natural gas as fuel. Some model-based approaches have been proposed in the fault analysis involving gas turbines ((Rasaienia, Moshiri and Moezzi, 2013), (Gupta *et al.*, 2008)) and two recent works ((Fontes and Pereira, 2016), (Fontes and Budman, 2017)) present clustering approaches applied to the same TPP (Figure 9).



(a) (b)

Figure 9 – (a) TPP and (b) Gas turbine RB211-G62 DF ((Fontes and Pereira, 2016), (Fontes and Budman, 2017), (Rolls-Royce, 2010))

The case study involved the recognition of fault patterns and normal operation during the starting of one of the three turbines in the Unit. The TPP has three gas turbines (GT), model Rolls Royce RB211-G62 DF, each one of these coupled to an electric generator in conjunction with other equipment to produce a total of 137 MW of electricity and 260.3 t/h steam.

One of the most frequent faults of the turbine comprises a high temperature difference between the combustion chamber sensors during the starting of this equipment (Trip by high temperature dispersion). The GT presented in Figure 9b has nine combustion chambers distributed radially around a central ring. There are 17 temperature sensors radially distributed around the combustion chambers. In order to protect the equipment from damage caused by differential expansion, the control system is designed to stop the turbine if the temperature of one sensor deviates by a value $\pm 150^{o}C$ from the average of the other temperature sensors. In this case, the interruption in the power supply leads to penalties applied to the TPP, causing financial loss. On the other hand, there is no a priori knowledge of a failure pattern (dynamic evolution of failure during start-up of the equipment) or even a clear definition of variables that could serve as indicators for its prediction.

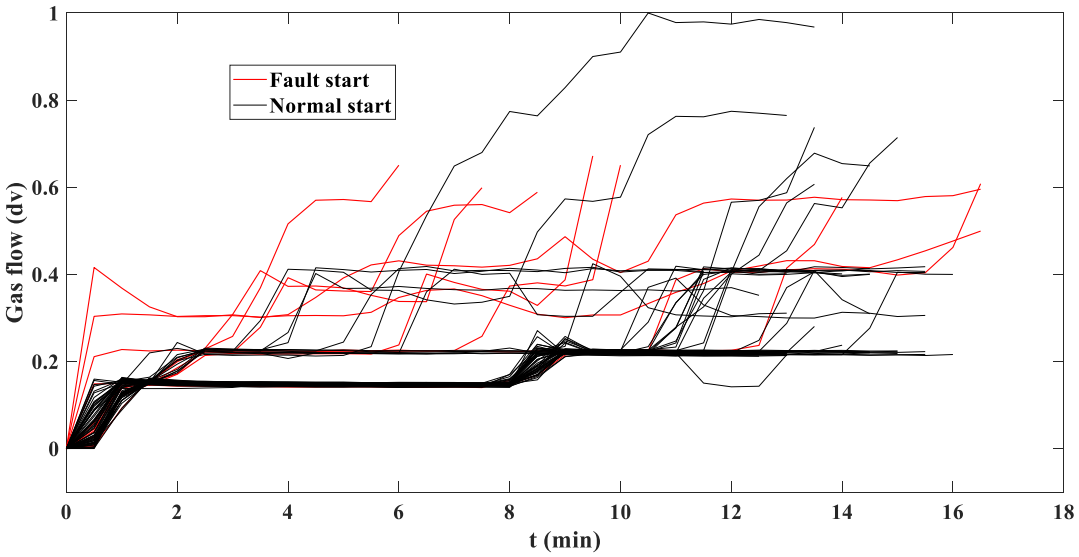
Two works ((Fontes and Budman, 2017), (Fontes and Pereira, 2016)) propose optimization-based clustering approaches, both inspired in the classical FCM, in order to classify and

distinguish between faulty and normal starts as well as recognize patterns that could be used as reference to support the development of a supervisory system for real time monitoring and prediction of faulty start-ups. Both works followed a MTS-based approach in which each object, extracted from the historical data base, involves three time series related to the three process variables (the flow of natural gas, F_g, the inlet temperature of the natural gas, T_i, and the temperature of the exhaust gas, T_e). Despite the good clustering and classification results, there was no analysis of the dynamic behavior of the process, the cause-effect relationship between these process variables and the feasibility (or achievability) of the recognized patterns.

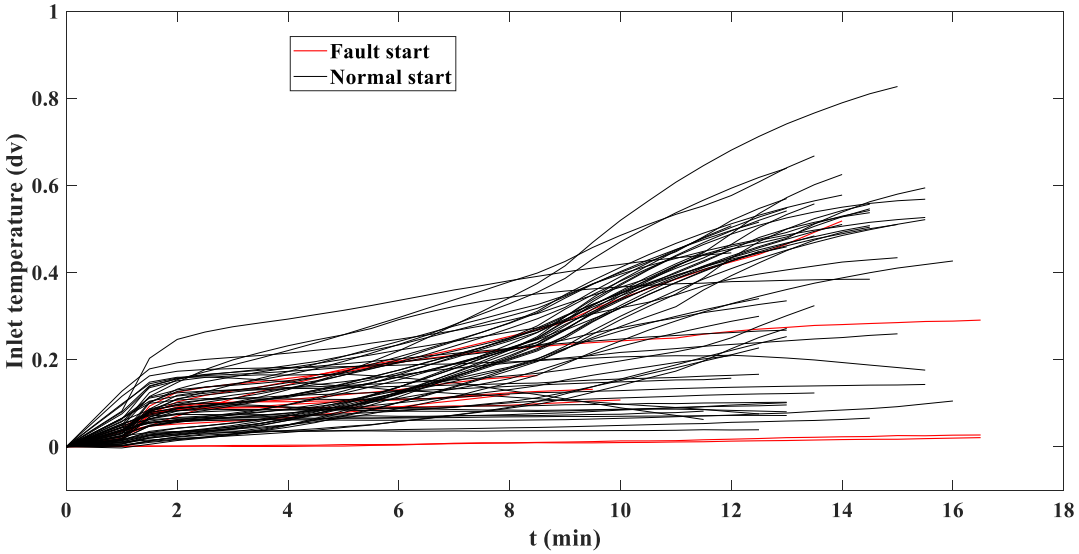
4.2.1 Data base modeling and results

70 objects (turbine starts) were available from the Plant Information Management System (PIMS) for the period between 2008 to 2011, consisting of an unbalanced data set with 60 objects associated to normal starts and 10 associated to starts with trip (failure). The sampling period was equal to 0.5 min and, as in previous section, each time series was normalized within the range [0;1] considering the maximum and minimum values of the respective process variable throughout the entire sample. Then each time series was also handled in Deviation Variable (dv) format.

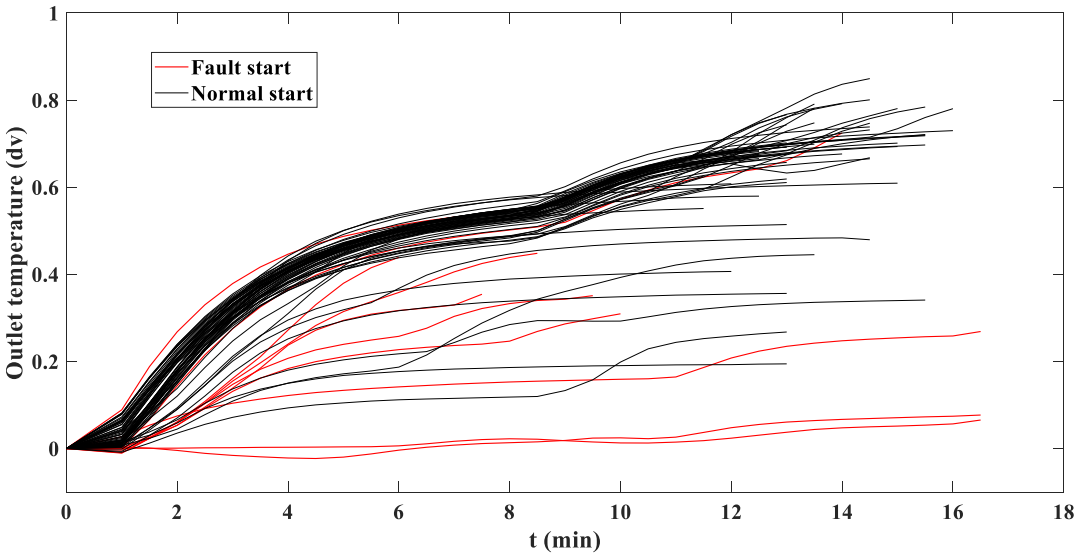
The clustering and pattern recognition (unsupervised learning) was carried out considering a training sample consisting of 10 trips and 30 randomly chosen normal objects. The remaining 30 normal objects were selected to compose a testing sample in order to cross-validate the quality of the recognized patterns with respect to the classification of test objects. Figure 10 presents all the time series that were used in the training sample. As can be seen, the length of time window associated with each MTS is not uniform because the starting time of the turbine (time interval until the exhausted gas temperature reaches 95% of its steady state value) is not always the same. However, the window associated to the pattern of each cluster was set equal to the maximum time interval ($\cong 16\ min$) verified in the sample (Figure 10) and therefore the application of SPCA (Eqs. 16 and 19) involved MTS with different lengths.


Fault start
Normal start
Gas flow (dv)
t (min)
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
14
16
18

(a)


Fault start
Normal start
Inlet temperature (dv)
t (min)
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
14
16
18

(b)



(c)

Figure 10 – Training sample. (a) Flow of natural gas; (b) Inlet temperature; (c) Outlet temperature.

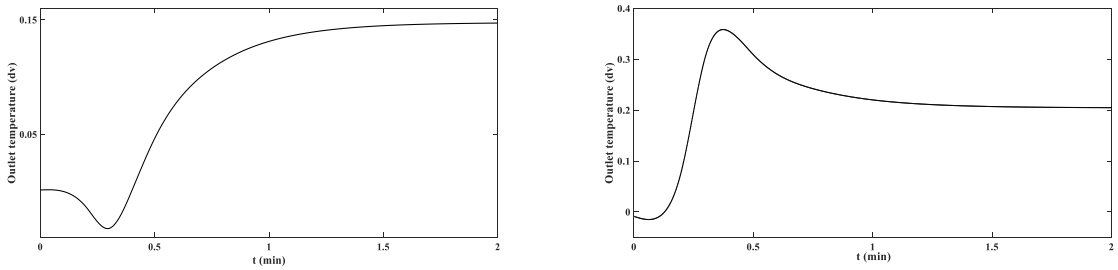
There is no clear understanding of the factors resulting in trips and/or unmeasured disturbances that really contribute to failed start-ups (trips). Furthermore, a phenomenological model capable of describing the dynamic behavior of the turbine is not available. On the other hand, each MTS involves two input variables (conditions of feeding the gas stream into turbine, F_g and T_i) and an output variable (outlet temperature of the natural gas, T_e) and, in this case, a reconciled pattern is one in which the profile or behavior of the gas outlet temperature is consistent with the changes in both inputs $(F_g$ and $T_i)$ (Prototype Inputs, PI) according to some dynamic model.

Using the same data sample (60 normal starts and 10 starts with trip), the following MISO (Multiple Input Single Output) ARX model with two inputs $(F_g$ and $T_i)$ and one output (T_e) was identified:

$$\hat{T}_e(k_t) - 0.699 \cdot \hat{T}_e(k_t - 1) = -0.266 \cdot F_g(k_t) + 0.487 \cdot F_g(k_t - 1) + 1.978 \cdot T_i(k_t) - 1.671 \cdot T_i(k_t - 1) + e(k_t) \quad (33)$$

$\hat{T}_e(k_t)$ is the predicted output at time k_t. Despite its apparent simplicity, the linear model (Eq. 33) is able to predict the dominant dynamic effect of the inputs $(F_g$ and $T_i)$ on the output (T_e) and is consistent with the expected features such as zero dead time (which should be

attributed to the small sampling period, 0.5 min) and positive static gain associated with both inputs (Figure 11).



(a) (b)

Figure 11 – Temperature of the exhaust gas (dv). Dynamic response (ARX model) to a step change in the gas flow (F_g) (a) and in the inlet temperature (T_i) (b).

The orders, parameters and dead times of each input in Eq. 33 were selected from a set of options in order to obtain the best fit with respect to the output measurements considering all objects. On the other hand, each object (MTS) comprises three time series extracted in the same period of time (same time window) but different objects are related to different time windows (different operating periods). Therefore, the data structure consists of a typical three-way array involving objects ("batches")×variables×time.

The overall sum of prediction errors in the entire sample was used as a metric to evaluate the quality of the model.

$$\sum_{i=1}^{70}\sum_{j=1}^{m_i}\left(\hat{T}_e(k_{tj}) - T_{ei}(k_{tj})\right)^2$$

m_i is the length of time window of the object i $(i = 1, \ldots, 70)$ and $\hat{T}_e(k_{tj})$ and $T_{ei}(k_{tj})$ are the predicted and measured output of the object i at time instant k_{tj}.

As in the previous case study (CSTR), this problem does not require the identification of specific models for the failure or normal operating conditions. Both states (failure and normal operation) are related to the same equipment or process (gas turbine) and the fault itself is just the effect of some disturbance (measurable, non-measurable or even unknown).

The results (percentage of misclassification and patterns) obtained with the simultaneous approach (Eqs. 27-32) (Case II, reconciled patterns) were compared with the FCM without pattern reconciliation (Case I, unreconciled patterns, $\alpha = 1$ in Eq. 27). Both approaches (Cases I and II) were able to gather in the same cluster 80% of the fault objects. This result, by itself, highlights the efficiency of the clustering procedure and the metric (SPCA) adopted

for the recognition of dissimilarities between objects, considering the unsupervised nature of learning and the availability of few trip data (minority class) in the original sample. Moreover, previous studies about the same turbine ((Fontes and Pereira, 2016), (Fontes and Budman, 2017)) show that two of the ten fault objects (20%) present in the training sample are significantly different from the other trip (fault) objects which suggests that the percentage of misclassification equal to 20% is really the best classification result obtained from the available sample. This can be attributed to the existence of a possible additional trip (failure) pattern that had not been identified due to the small amount of data/information related to the fault state.

In both cases (I and II), the best classification results (Table 3) were obtained considering only two clusters. The fault and normal cluster have 80% of the fault and 87-90 % of the normal objects, respectively, present in the entire training sample. The recognized patterns of each cluster were used as reference for classification (according to similarity) of the test sample objects (30 normal startups). In both cases only one object of the test sample was classified as belonging to the fault cluster (3.3% misclassification) which shows a good cross-validation result.

Table 3 – Percentage of misclassifications

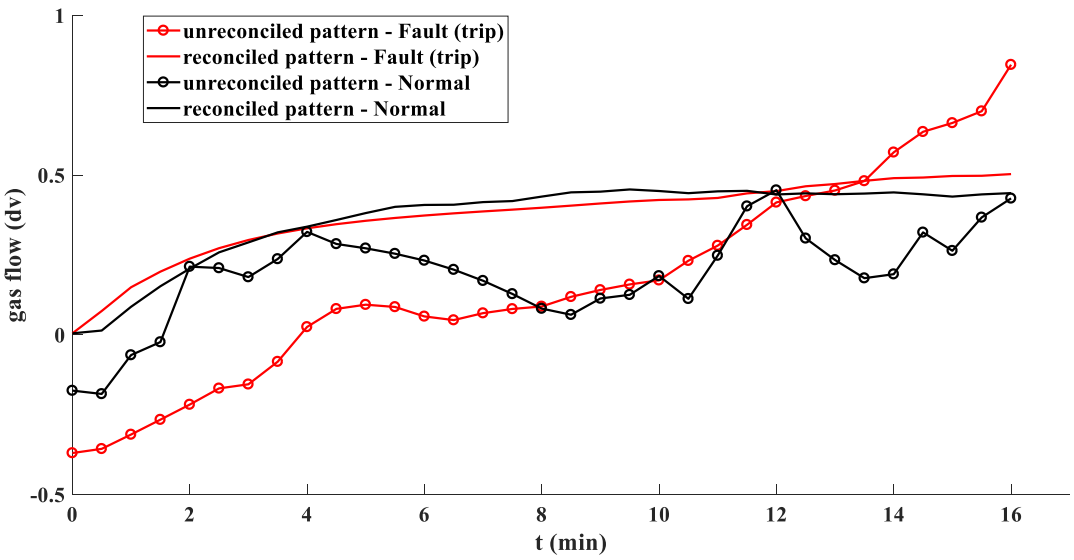
	Case I (unreconciled patterns)	Case II (reconciled patterns)
Training data (10 trip and 30 normal objects)	F – 20% N – 10%	F – 20% N – 13%
Test data (30 normal objects)	N – 3.3%	N – 3.3%

F – trip/fault objects; N – normal objects

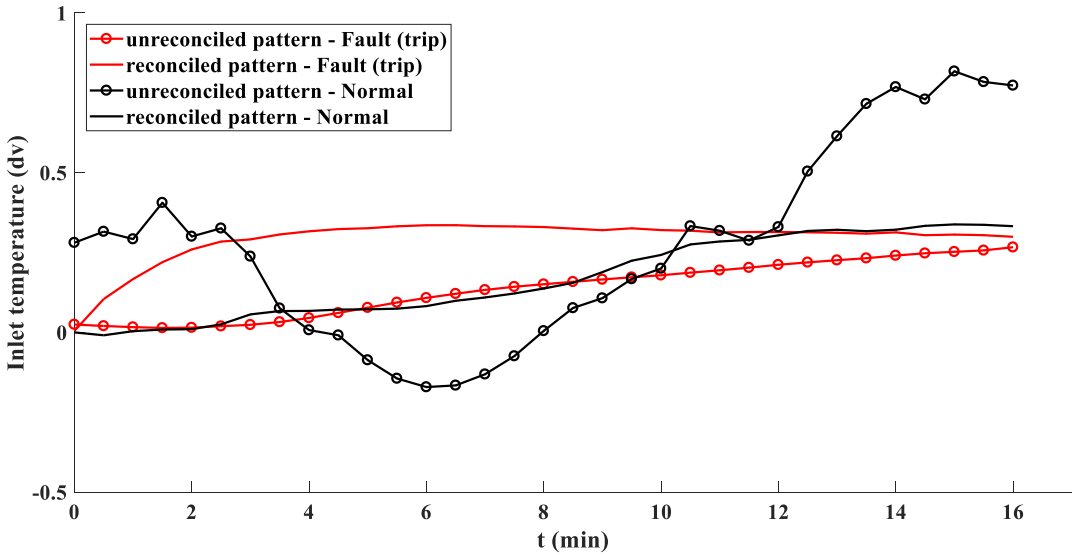
The results presented in Table 3 highlight two aspects already verified in the previous case study. First, the soft (Eq. 27) and hard constraints (Eqs. 31 and 32) associated with the clustering problem with pattern reconciliation do not worsen the classification results and the quality of the clustering. Second, even unreconciled patterns are able to achieve good clustering and classification results.

Figure 12 presents the patterns recognized in Cases I and II. The simultaneous reconciliation approach does not only change the output profile $\left(T_e(t)\right)$ but also the inputs

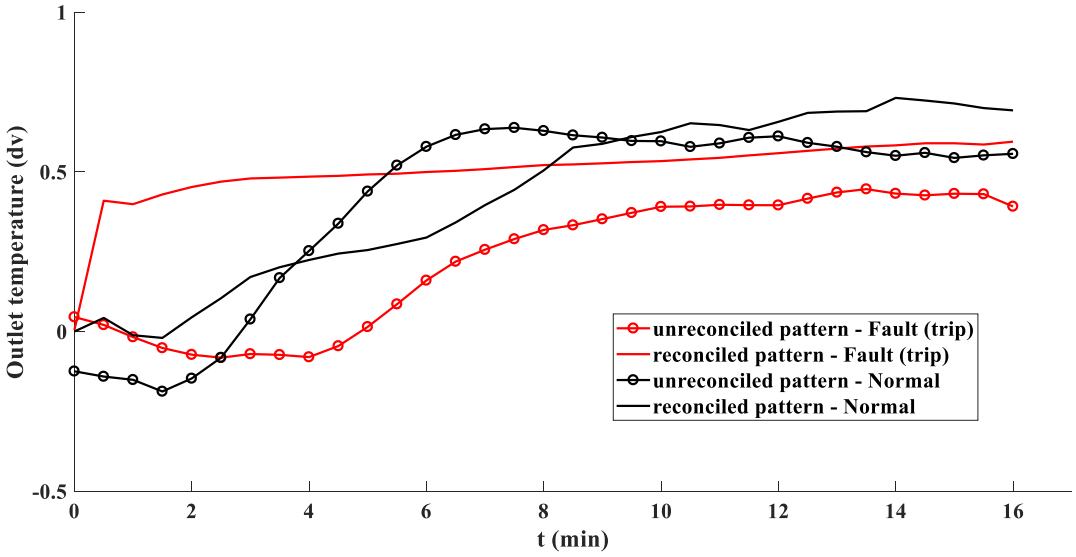
$\left(F_g(t) \text{ and } T_i(t)\right)$ since these are also part of each prototype. Despite the small differences between the dynamics of the gas flow in both reconciled patterns (fault and normal, Figure 12a), the profiles related to the inlet temperatures are less similar and justify the differences verified in the output (outlet temperature) (Figures 12b-c). Moreover, the differences between reconciled patterns (fault and normal) are more pronounced in the dynamic behavior of the inlet and outlet temperatures which are sufficient to generate changes in the directions of the principal components capable of recognizing dissimilarities among fault and normal objects. Figure 13 presents the output (outlet temperature) profiles predicted by the dynamic model (Eq. 33) considering the same input profiles $\left(F_g(t) \text{ and } T_i(t)\right)$ of each pattern. Figure 13b shows significant differences between the unreconciled output pattern (normal cluster) and the dominant dynamics predicted for the process. Figure 13a shows the similarity between the reconciled output patterns (both clusters) and the expected dynamic for the process.



(a)

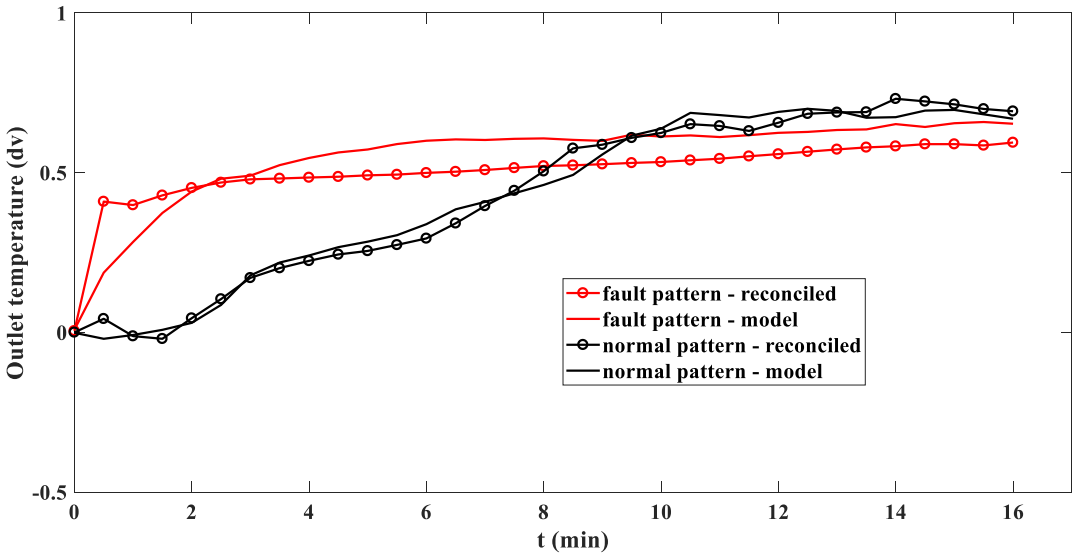


(b)

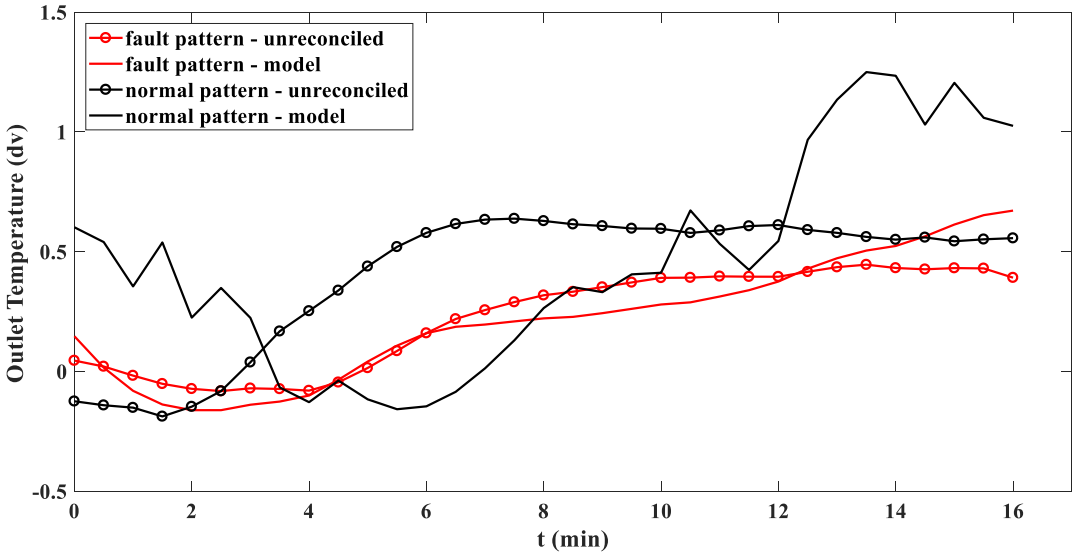


(c)

Figure 12 – Reconciled and unreconciled Patterns. (a) Flow of natural gas; (b) Inlet temperature; (c) Outlet temperature.



(a)



(b)

Figure 13 – Outlet temperature ($T_e(t)$) – patterns recognized and predicted by the process model. (a) reconciled and (b) unreconciled.

5. Conclusions

A novel approach was proposed for the adjustment (reconciliation) of patterns to the dynamics of a process in the clustering and classification of multivariate time series. The strategy is based on a classic Fuzzy C-Means (FCM) algorithm and considers the process dynamics as a soft constraint in order to ensure the feasibility of the recognized patterns. The

problem consists of bi-criterion optimization subject to hard constraints associated with the process variables, process model and membership degrees.

In problems involving fault detection, patterns related to the normal and/or failure condition, identified/recognized through a clustering procedure, can be used as reference to predict correct trajectories for normal operation or to support the development of a supervisory system for real time monitoring, diagnosis and fault prediction. In such cases, it is important that the prototypes/patterns (or references) be achievable or consistent with the dominant dynamic behavior of the process by presenting, among other things, coherent signals of static gains between inputs and outputs.

This paper shows that only clustering involving MTS does not ensure the coherence of the patterns in relation to the process. Furthermore, the recognition of feasible patterns would be even more difficult in cases where the data sample has little information and there is low homogeneity among objects of the same class (same label). This is a typical and frequent situation in data extracted from real industrial systems subject to noise and unknown disturbances.

This work presents a systematic way to cope with the clustering and reconciliation of patterns involving MTS, according to the type of process variables (input or output) considered in each object. The problem of pattern reconciliation is categorized into two sub-problems by defining and distinguishing between prototype and non-prototype inputs (PI and NPI) and, for each type of problem, a generic model of constrained optimization and the strategy of resolution (simultaneous or sequential) are proposed. This work therefore presents a feasible alternative to include background knowledge related to the dynamic behavior of the process to guide a clustering and pattern recognition algorithm involving multivariate time series.

Two applications are presented: a simulation case study involving the availability of a phenomenological model of the process (CSTR) and a real industrial case, based on historical data, involving an empirical dynamic model. The applications comprise the fault diagnosis (CSTR) and detection (gas turbine). The results show that the strategies proposed are able to obtain a reconciled pattern without compromising the quality of the clustering and classification results, i.e. without impairing the ability to diagnose or detect failures. Therefore, the proposed method provides a satisfactory classification model of objects from achievable patterns.

In addition, this paper presents the analytical solution for the reconciliation problem involving Univariate Time Series (UTS) and shows that in such cases meaningful differences between unreconciled and reconciled patterns are not expected.

Although this work is based on the use of SPCA as a similarity metric, the proposed reconciliation approach is not limited to a specific metric. Other similarity metrics can be employed without altering the structure of the proposed models and concepts.

Acknowledgements

The authors acknowledge the financial support provided by the Federal Agency for Support and Evaluation of Graduate Education (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES-BRAZIL) and the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq-BRAZIL).

Appendix

Applying the first order condition to a given reconciled pattern in the optimization problem defined by the equations 6-12, we have:

$$\frac{\partial H_\varepsilon}{\partial v_i^r} = \left[\sum_{k=1}^{n}\left(u_{ik}^{\varepsilon}\cdot\frac{\partial\left(\|x_k - v_i^r\|^2\right)}{\partial v_i^r}\right)\right] + \left[\frac{\partial\left(\|v_i^r - \hat{y}_{c_i}\|^2\right)}{\partial v_i^r}\right] = 0 \qquad (A.1)$$

where

$$\left[\frac{\partial\left(\|v_i^r - \hat{y}_{c_i}\|^2\right)}{\partial v_i^r}\right] = \lim_{\gamma\to 0}\frac{\|(v_i^r + \gamma\cdot\xi) - \hat{y}_{c_i}\|^2 - \|v_i^r - \hat{y}_{c_i}\|^2}{\gamma}, \quad \gamma \in \Re \text{ and } \xi \in \Re^m$$

Considering that $\|\quad\|$ refers to the Euclidean distance,

$$\left[\frac{\partial\left(\|v_i^r - \hat{y}_{c_i}\|^2\right)}{\partial v_i^r}\right]$$

$$= \lim_{\gamma\to 0}\frac{\left\{\left(v_i^r + \gamma\cdot\xi - \hat{y}_{c_i}\right)^T\cdot\left(v_i^r + \gamma\cdot\xi - \hat{y}_{c_i}\right) - \left(v_i^r - \hat{y}_{c_i}\right)^T\cdot\left(v_i^r - \hat{y}_{c_i}\right)\right\}}{\gamma}$$

$$\left[\frac{\partial\left(\|v_i^r - \hat{y}_{c_i}\|^2\right)}{\partial v_i^r}\right]$$

$$= \lim_{\gamma\to 0}\frac{\left\{\left[\left(v_i^r - \hat{y}_{c_i}\right)^T + \gamma\cdot\xi^T\right]\cdot\left[\left(v_i^r - \hat{y}_{c_i}\right) + \gamma\cdot\xi\right] - \left(v_i^r - \hat{y}_{c_i}\right)^T\cdot\left(v_i^r - \hat{y}_{c_i}\right)\right\}}{\gamma}$$

$$\left[\frac{\partial\left(\left\|v_i^r-\hat{y}_{c_i}\right\|^2\right)}{\partial v_i^r}\right]$$

$$=\lim_{\gamma\to 0}\frac{\left[\left(v_i^r-\hat{y}_{c_i}\right)^T\cdot\left(v_i^r-\hat{y}_{c_i}\right)-\left(v_i^r-\hat{y}_{c_i}\right)^T\cdot\gamma\cdot\xi-\gamma\cdot\xi^T\cdot\left(v_i^r-\hat{y}_{c_i}\right)+\gamma^2\cdot\xi^T\cdot\xi-\left(v_i^r-\hat{y}_{c_i}\right)^T\cdot\left(v_i^r-\hat{y}_{c_i}\right)\right]}{\gamma}$$

which leads to

$$\left[\frac{\partial\left(\left\|v_i^r-\hat{y}_{c_i}\right\|^2\right)}{\partial v_i^r}\right]=\lim_{\gamma\to 0}\frac{\left[2\cdot\gamma\cdot\left(v_i^r-\hat{y}_{c_i}\right)^T\cdot\xi+\gamma^2\cdot\xi^T\cdot\xi\right]}{\gamma}\cong 2\cdot\left(v_i^r-\hat{y}_{c_i}\right)^T\cdot\xi \qquad \text{(A.2)}$$

since $\gamma^2\cdot\xi^T\cdot\xi\to 0$

Analogously,

$$\sum_{k=1}^n\left(u_{ik}^{\varepsilon}\cdot\frac{\partial\left(\left\|x_k-v_i^r\right\|^2\right)}{\partial v_i^r}\right)=-2\cdot\sum_{k=1}^n u_{ik}^{\varepsilon}\cdot\left(x_k-v_i^r\right)^T\cdot\xi \qquad \text{(A.3)}$$

Using (A.1), (A.2) and (A.3):

$$-\left(\sum_{k=1}^n u_{ik}^{\varepsilon}\cdot\left(x_k-v_i^r\right)^T\right)+\left(v_i^r-\hat{y}_{c_i}\right)^T=0$$

$$-\left(\sum_{k=1}^n u_{ik}^{\varepsilon}\cdot\left(x_k-v_i^r\right)\right)+\left(v_i^r-\hat{y}_{c_i}\right)=0$$

$$v_i^r\cdot\left(\sum_{k=1}^n u_{ik}^{\varepsilon}\right)+\left(v_i^r-\hat{y}_{c_i}\right)=\sum_{k=1}^n u_{ik}^{\varepsilon}x_k$$

$$v_i^r=\frac{\sum_{k=1}^n u_{ik}^{\varepsilon}x_k}{\left(\sum_{k=1}^n u_{ik}^{\varepsilon}\right)+1}+\frac{\hat{y}_{c_i}}{\sum_{k=1}^n u_{ik}^{\varepsilon}}$$

or $v_i^r\cong v_i+\frac{\hat{y}_{c_i}}{\sum_{k=1}^n(u_{ik})^{\varepsilon}}$, where $v_i=\frac{\sum_{k=1}^n(u_{ik})^{\varepsilon}\cdot x_k}{\sum_{k=1}^n(u_{ik})^{\varepsilon}}$. (A.4)

References

Aghabozorgi, S., Shirkhorshid, A. S. and Wah, T. Y. (2015) "Time-series clustering – A decade review," *Information Systems*, 53, pp. 16–38.

Bankó, Z. and Abonyi, J. (2012) "Correlation based dynamic time warping of multivariate time series," *Expert Systems with Applications*, 39, pp. 12814–12823.

Bezdek, J. C. *et al.* (2005) *Fuzzy Models and Algorithms for pattern recognition and image processing*. New York: Springer Science+Business Media, Inc.

CAO, H. *et al.* (2012) "Integrated Oversampling for Imbalanced Time Series Classification," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 25(12), pp. 2809–2822.

Charikar, M., Guruswami, V. and Wirth, A. (2005) "Clustering with qualitative information," *Journal of Computer and System Sciences*, 71(360–383).

Coppi, R., D'urso, P. and Giordani, P. (2010) "A Fuzzy Clustering Model for Multivariate Spatial Time Series,"

Journal of classification, 27, pp. 54–88.

D'urso, P. (2004) "Fuzzy C-Means Clustering Models for Multivariate Time-Varying Data: Different Approaches," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12(3), pp. 287–326.

Dao, T.-B.-H., Duong, K.-C. and Christel Vrain (2017) "Constrained clustering by constraint programmin," *Artificial Intelligence*, 244, pp. 70–94.

Deng, X., Tian, X. and Chen, S. (2013) "Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis," *Chemometrics and Intelligent Laboratory Systems*, 127, pp. 195–209.

Diez-Olivan, A. *et al.* (2017) "Data-driven prognostics using a combination of constrained K-means clustering, fuzzy modeling and LOF-based score," *Neurocomputing*, 241(97–107).

Dobos, L. and Abonyi, J. (2012) "On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segmentation," *Chemical Engineering Science2*, 75, pp. 96–105.

Fontes, C. H. and Budman, H. (2017) "A Hybrid Clustering Approach for Multivariate Time Series - A Case Study applied to failure analysis in a Gas Turbine," *ISA Transactions*, in press.

Fontes, C. H. and Pereira, O. (2016) "Pattern recognition in multivariate time series - A case study applied to fault detection in a gas turbine," *Engineering Applications of Artificial Intelligence*, 49, pp. 10–18. doi: 10.1016/j.engappai.2015.11.005.

Fu, T. (2011) "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, 24, pp. 164–181.

Ganganwar, V. (2012) "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, 2(4), pp. 42–47.

Grossi, V., Romei, A. and Turini, F. (2017) "Survey on using constraints in data mining," *Data Mining and Knowledge Discovery*, 31, pp. 424–464.

Gupta, S. *et al.* (2008) "Fault detection and isolation in aircraft gas turbine engines. Part 1: Underlying concept," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 222(3), pp. 307–318.

Harrou, F. *et al.* (2015) "Improved principal component analysis for anomaly detection: Application to an emergency department," *Computers & Industrial Engineering*, 88, pp. 63–77.

Hoppner, F. *et al.* (1999) *Fuzzy Cluster Analysis – Methods for Classification, Data Anaysis and Image Recognition*. New York: John Wiley & Sons.

Izakian, H., Pedrycz, W. and Jamal, I. (2015) "Fuzzy clustering of time series data using dynamic time warping

distance," *Engineering Applications of Artificial Intelligence*, 39, pp. 235–244.

K. Wagstaff *et al.* (2001) "Constrained k-means clustering with background knowledge Learning," in *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, USA: Morgan Kaufmann Publishers Inc., pp. 577–584.

Kavitha, V. and Punithavalli, M. (2010) "Clustering Time Series Data Stream - A Literature Survey," *International Journal of Computer Science and Information Security*, 8(1), pp. 289–294.

Keogh, E. J. and Kasetty, S. (2002) "On the need for time series data mining benchmarks: a survey and empirical demonstration," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*. Edmonton (Alberta) - Canada, pp. 23–26.

Khediri, I. Ben, Limam, M. and Weihs, C. (2011) "Variable window adaptive Kernel Principal Component Analysis for nonlinear nonstationary process monitoring," *Computers & Industrial Engineering*, 61(3), pp. 437–446.

Kiri Lou Wagstaff (2002) *Intelligent Clustering with Instance-Level Constraints*. Faculty of the Graduate School of Cornell University.

Lampert, T. *et al.* (2018) "Constrained distance based clustering for time-series: a comparative and experimental study," *Data Mining and Knowledge Discovery*. doi: https://doi.org/10.1007/s10618-018-0573-y.

Law, M. H. C., Topchy, A. and Jain, A. K. (2004) "Clustering with soft and group constraints," in *Proceedings of the Joint IAPR International Workshops on Structural, Syntactic,and Statistical Pattern Recognition*, pp. 662–670.

Li, X. *et al.* (2018) "A data-driven model for the air-cooling condenser of thermal power plants based on data reconciliation and support vector regression," *Applied Thermal Engineering*, 129, pp. 1496–1507.

Liao, T. W. (2005) "Clustering of time series data - a survey," *Pattern Recognition*, 38, pp. 1857 – 1874.

Mitsa, T. (2010) *Temporal Data Mining*. CRC Press, Taylor & Francis Group.

Oliveira, R. M. de, Chaves, A. A. and Lorena, L. A. N. (2017) "A comparison of two hybrid methods for constrained clustering problems," *Applied Soft Computing*, 54(256–266).

Petitjean, F., Ketterlin, A. and Gancarski, P. (2011) "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, 44(3), pp. 678–693.

Plant, C., Wohlschlager, A. M. and Zherdin, A. (2009) "Interaction-based Clustering of Multivariate Time Series," in *Proceedings of the Ninth IEEE International Conference on Data Mining*. Miami- Floria (USA), pp. 914–919.

Rasaienia, A., Moshiri, B. and Moezzi, M. (2013) "Mechanical Systems and Signal Processing," *Turkish Journal of Electrical Engineering and Computer Sciences*, 21(5), pp. 1340–1350.

Rolls-Royce (2010) “Training Manual RB 211-G62 DF.”

Seret, A., Verbraken, T. and Baesens, B. (2014) “A new knowledge-based constrained clustering approach: Theoryand application in direct marketing,” *Applied Soft Computing*, 24, pp. 316–327.

Singhal, A. and Seborg, D. E. (2002) “Pattern Matching in Multivariate Time Series Databases Using a Moving-Window Approach,” *Industrial and Engineering Chemistry Research*, 41, pp. 3822–3838.

Singhal, A. and Seborg, D. E. (2006) “Evaluation of a pattern matching method for the Tennessee Eastman challenge process,” *Journal of Process Control*, 16, pp. 601–613.

Syed, M. S. *et al.* (2016) “Enhanced turbine monitoring using emissions measurements and data reconciliation,” *Applied Energy*, 173, pp. 355–365.

Trebuňa, P. and Halčinová, J. (2013) “Mathematical Tools of Cluster Analysis,” *Applied Mathematics*, 4, pp. 814–816.

Vaidyanathan, R. and Venkatasubramanian, V. (1992) “Representing and Diagnosing Dynamic Process Data Using Neural Networks,” *Engineering Applications of Artificial Intelligence*, 5, pp. 11–21.

Venkatasubramanian, V. *et al.* (2003) “A review of process fault detection and diagnosis - Part I: Quantitative model-based methods,” *Computers and Chemical Engineering*, 27, pp. 293–311.

Wang, X. *et al.* (2013) “Experimental comparison of representation methods and distance measures for time series data,” *Data Mining and Knowledge Discovery*, 26, pp. 275–309.

Wu, D., Kwon, R. H. and Costa, G. (2017) “A constrained cluster-based approach for tracking the S&P 500 index,” *International Journal of Production Economics*, 193(222–243).

Xun, L. and Zhishu, L. (2010) “The similarity of multivariate time series and its application,” in *Proceedings of the International Conference on Management of e-Commerce and e-Government, Sichuan, China*, pp. 76–81.

Yang, K. and Shahabi, C. (2004) “A PCA-based Similarity Measure for Multivariate Time Series,” in *Proceedings of the International Workshop on Multimedia Databases, ACM-MMDB, Washington DC, USA*, pp. 1–10.

Yang, Y. and Jianmin Jiang (2018) “Bi-weighted ensemble via HMM-based approaches for temporal data clustering,” *Pattern Recognition*, 76, pp. 391–403.

Zakaria, J. *et al.* (2016) “Accelerating the discovery of unsupervised-shapelets,” *Data Mining and Knowledge Discovery*, 30, pp. 243–281.

UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA

PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA INDUSTRIAL - PEI

Rua Aristides Novis, 02, 6º andar, Federação, Salvador BA
CEP: 40.210-630
Telefone: (71) 3283-9800
E-mail: pei@ufba.br
Home page: http://www.pei.ufba.br