



UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA
SIMULAÇÃO E OTIMIZAÇÃO DE PROCESSOS

DETECÇÃO DE FALHAS EM PROCESSOS QUÍMICOS
CONTÍNUOS: UMA ABORDAGEM VIA *ENSEMBLE*
***LEARNING* E INFERÊNCIA BAYESIANA**

Douglas Fernandes Rodrigues da Silva
Orientador: Prof. Dr. Gustavo Matheus de Almeida

Belo Horizonte– MG – Brasil

2019



UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA
SIMULAÇÃO E OTIMIZAÇÃO DE PROCESSOS

DETECÇÃO DE FALHAS EM PROCESSOS QUÍMICOS
CONTÍNUOS: UMA ABORDAGEM VIA *ENSEMBLE*
***LEARNING* E INFERÊNCIA BAYESIANA**

Douglas Fernandes Rodrigues da Silva

Trabalho de qualificação apresentado ao Programa de Pós-Graduação em Engenharia Química como parte dos requisitos necessários à obtenção do título de Mestre em **Engenharia Química**.

Belo Horizonte – MG – Brasil

2019

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me deu a oportunidade de concluir o curso de Mestrado em Engenharia Química na Universidade Federal de Minas Gerais. À minha mãe e familiares pelo apoio, pois sem eles não conseguiria esse passo importante na minha vida.

Agradeço ao meu orientador Gustavo Matheus de Almeida, que me atendeu prontamente, pela paciência, pelo grande ensinamento e pela amizade.

Agradeço também a toda estrutura da Engenharia Química da UFMG pelo apoio e incentivo à produção do trabalho, e a CAPES por ter acreditado e me financiado como bolsista.

Enfim, a todos aqueles que, em meio às dificuldades, de alguma forma, contribuíram para o meu crescimento e apoio durante este tempo de trabalho.

SUMÁRIO

1. INTRODUÇÃO	1
2. OBJETIVOS	3
2.1. Objetivo Geral	3
2.2. Objetivos Específicos	4
3. REVISÃO BIBLIOGRÁFICA	5
3.1. Histórico sobre Controle de Processos	5
3.2. Modelos Fenomenológicos x Modelos baseados em Dados	6
3.3. Monitoramento de Processos.....	7
3.4. Carta de Controle.....	9
3.5. Controle Estatístico de Processos(CEP), Univariado	10
3.6. Controle Estatístico de Processos (CEP), Multivariado	11
4. ANÁLISE POR COMPONENTES PRINCIPAIS (PCA)	14
4.1. Análise por Componentes Principais via <i>Kernel</i> (KPCA).....	17
4.2. Análise por Componentes Principais, Dinâmica (DPCA).....	18
4.3. Combinação entre Técnicas	19
5. BENCHMARK TENNESSEE	23
6. METODOLOGIA	26
6.1. Etapa 1: Construção e Validação dos Sistemas de Detecção de Falhas.....	26
6.2. Etapa 2: Combinação dos Sistemas de Detecção de Falhas.....	27
7. RESULTADOS E DISCUSSÃO	28
7.1. Etapa 1: Construção e Validação dos Sistemas de Detecção de Falhas.....	28
7.2. Etapa 2: Combinação dos Sistemas de Detecção de Falhas.....	40
8. CONCLUSÕES	51
8.1. Sugestões de Trabalhos Futuros	52
9. REFERÊNCIAS BIBLIOGRÁFICAS	53

LISTA DE FIGURAS

Figura 1: Carta de controle (a) fora e (b) sob controle estatístico (Control, 1986).....	10
Figura 2: (a) Cartas de controle univariadas para as variáveis X_1 e X_2 , e (b) região de controle definida por ambas (Montgomery, 2008).....	12
Figura 3: Mapeamento entre o espaço de entrada (RN) e o espaço de características (F) através de uma função de <i>kernel</i> (φ), em que x é o vetor de observações originais (Scholkopf et al., 1999).....	18
Figura 4: Fluxograma de processo do <i>benchmark</i> Tennessee (Downs e Vogel, 1993).....	24
Figura 5: Variância explicada acumulada (%).	32
Figura 6: Valores T^2 e Q para as falhas 4 e 5, dado o modelo PCA.	35
Figura 7: Valores T^2 e Q para as falhas 4 e 5, dado o modelo DPCA.....	37
Figura 8: MDR (<i>Missed Detection Rate</i>), relativo às estatísticas (a) T^2 e (b) Q , para os sistemas individuais (PCA, DPCA e KPCA) e aquele combinado (EL).	45
Figura 9: Resultados relativos à estatística T^2 , em que 1 refere-se à falha 1, 2 à falha 2, e assim por diante até a falha 21, e 0, à operação normal.....	48
Figura 10: Resultados relativos à estatística Q , em que 1 refere-se à falha 1, 2 à falha 2, e assim por diante até a falha 21, e 0, à operação normal.....	50

LISTA DE TABELAS

Tabela 1: Conjunto de falhas no <i>benchmark</i> Tennessee (Russell et al., 2000; Dows e Vogel, 1993).	25
Tabela 2: Autovalores em ordem decrescente para o modelo PCA.	29
Tabela 3: Autovalores em ordem decrescente para o primeiro modelo KPCA, usado para fins de validação.....	29
Tabela 4: Autovalores em ordem decrescente para o modelo DPCA.....	31
Tabela 5: Limites superior de controle para T^2 e Q.	33
Tabela 6: Modelo PCA: Validação via comparação com Russell et al. (2000) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, <i>Missed Detection Rate</i>).	38
Tabela 7: Modelo DPCA: Validação via comparação com Russell et al. (2000) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, <i>Missed Detection Rate</i>).	39
Tabela 8: Método KPCA: Validação via comparação com Ge e Song (2013) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, <i>Missed Detection Rate</i>).	40
Tabela 9: Autovalores em ordem decrescente para o segundo modelo KPCA, usado para fins de construção do sistema de detecção de falhas.	41
Tabela 10: Limites superior de controle para T^2 e Q.	41
Tabela 11: Taxa de detecções perdidas (MDR, <i>Missed Detection Rate</i>), relativas à métrica de monitoramento T^2 , para os modelos individuais (PCA, DPCA e KPCA) e aquele combinado (EL).	44
Tabela 12: Taxa de detecções perdidas (MDR, <i>Missed Detection Rate</i>), relativas à métrica de monitoramento Q, para os modelos individuais (PCA, DPCA e KPCA) e aquele combinado (EL).	45

LISTA DE ABREVIATÖES

PCA–Análise por Componentes Principais (*Principal Component Analysis*)

KPCA– Análise por Componentes Principais via *Kernel* (*Kernel PCA*)

DPCA–Análise por Componentes Principais, Dinâmica (*Dynamic PCA*)

T^2 –Estatística de Hotelling

Q –Erro de Predição Quadrático (SPE; *Squared Prediction Error*)

MDR – Taxa de detecções perdidas (*Missed Detection Rate*)

RESUMO

Uma falha é um evento indesejado em qualquer operação em uma indústria química. A sua ocorrência pode causar perdas em segurança operacional, ambiental e econômica. Devido à complexidade dos processos químicos por um lado, e a disponibilidade de grandes massas de dados por outro, atualmente, é crescente o número de aplicações de sistemas de detecção de falhas baseados em dados históricos de processos. A Análise por Componentes Principais (PCA) é o principal método estatístico multivariado empregado com esse propósito. Porém, nenhuma técnica em particular é capaz de descrever qualquer processo químico por completo; desse modo, uma combinação de técnicas constitui-se em uma solução alternativa potencial para uma detecção de falhas mais eficiente. O objetivo do presente trabalho é investigar o uso combinado (*ensemble learning*) de três técnicas: PCA, *Kernel PCA* e PCA Dinâmico, através de uma abordagem via inferência Bayesiana. O estudo de caso é o benchmark Tennessee, o mais usual na comunidade de Engenharia Química em todo o mundo. Verificou-se, de modo quantitativo, a maior eficiência da abordagem baseada em *ensemble learning* via inferência Bayesiana, com o uso da métrica de monitoramento, taxa de detecções perdidas (*Missed Detection Rate*; MDR). A partir da análise dos resultados, observou-se que não há perda da capacidade de detecção do sistema combinado em relação àqueles individuais; a melhora significativa do desempenho, a partir da métrica taxa de detecções perdidas (*Missed Detection Rate*, MDR), principalmente para cinco de seis falhas de difícil detecção; com valores de MDR consideravelmente próximos a zero para quatro dessas falhas. Em resumo, foi possível verificar o ganho significativo ao se adotar um sistema combinado de detecção de falhas em relação aos modelos individuais. A maior certeza sobre a presença de uma falha é um fator crucial para pelo menos mitigar perdas em potencial em processos industriais.

Palavras-chave:

Detecção de falhas, Análise por Componentes Principais (PCA); *Kernel PCA* (KPCA); PCA Dinâmico (DPCA); *Ensemble learning*, Inferência bayesiana, *Benchmark Tennessee*.

ABSTRACT

A fault is an unwanted event in any operation in a chemical industry. Its occurrence may cause safety, environmental and economic losses. Due to the complexity of chemical processes from one side, and the availability of very large data sets from the other side, the use of fault detection systems based on historical process data has increased. The Principal Component Analysis (PCA) is the main multivariate statistical method employed with this purpose. However, no particular technique is capable of describing any chemical process as a whole; thus, the combination of techniques constitutes in a potential alternative solution towards more efficient fault detection. The objective of this work is to investigate the combined use (ensemble learning) of three techniques, namely PCA, kernel PCA and dynamic PCA, through a Bayesian inference strategy. The case study is the Tennessee benchmark, which is the most usual in the Chemical Engineering community worldwide. The higher efficiency of the adopted approach was quantitatively verified by means of the monitoring metric called Missed Detection Rate (MDR). From the analysis of the results, it was observed that there is no loss of the detection capacity of the combined system in relation to the individual ones; the significant improvement in performance, from the metric Missed Detection Rate (MDR), mainly for five of six hard-to-detect faults; with MDR values close to zero for four of these faults. In summary, it was possible to verify the significant gain when adopting a combined fault detection system in relation to the individual models. Greater certainty about the presence of a fault is a crucial factor in at least mitigating potential losses in industrial processes.

Keywords:

Fault Detection, Principal Component Analysis (PCA); Kernel PCA (KPCA); Dynamic PCA (DPCA); Ensemble learning; Bayesian inference; Benchmark Tennessee.

1. INTRODUÇÃO

A detecção de falhas é uma área com relevância histórica, por lidar com condições anormais dos processos químicos industriais em geral. Uma falha é um desvio de qualquer parâmetro ou variável em relação ao ponto normal de operação. Uma detecção precoce de falhas ainda pode evitar distúrbios em outras etapas do processo, e então reduzir perdas de produtividade em geral. Esse fator, além das questões de segurança operacional, produção limpa e econômica, é a razão para o contínuo interesse dos meios tanto acadêmico quanto industrial em detecção de falhas, com uma vasta literatura disponível a respeito (Severson et al., 2016).

Do ponto de vista da construção de sistemas de detecção de falhas, tem-se o uso de modelos fenomenológicos ou baseados em dados históricos, ou híbridos, uma combinação entre ambos. Os modelos fenomenológicos são baseados em métodos analíticos, com uma descrição dos mecanismos físico-químicos dos processos; desse modo, a abordagem é geralmente limitada a processos com poucas variáveis e conhecimento suficiente. Os modelos históricos são baseados em correlações construídas a partir de dados históricos sobre as operações dos processos. Esse fato explica o seu uso crescente em aplicações industriais, onde o conhecimento sobre as operações é geralmente insuficiente e a disponibilidade de dados é geralmente alta. As técnicas usuais nesse caso são aquelas de inteligência computacional, como rede neural artificial, e de estatística multivariada, como Análise por Componentes Principais (PCA; usado como ponto de partida neste trabalho). Em relação aos métodos estatísticos, PCA é a técnica mais comumente empregada em monitoramento de processos químicos, desde os anos 90. A abordagem híbrida é uma junção entre ambos os modelos (Yin et al., 2014; Das et al., 2012; Venkatasubramanian et al., 2003a,b,c; Chiang et al., 2001).

Um movimento crescente em todo o mundo, atualmente, é aquele denominado de Indústria 4.0. Este conceito é oriundo do avanço contínuo de inovações tecnológicas nas áreas de automação, controle e tecnologia da informação. Essa nova abordagem tem sido responsável por uma nova era industrial, denominada de fábrica inteligente, com impactos em toda a cadeia

de valor do processo produtivo. Além dos pilares de Internet das Coisas e de Sistemas ciber-físicos, outra base fundamental para a operacionalização das indústrias nesse contexto é a de Análise de dados. Esse ramo visa a uma tomada de decisão mais racional, onde também se destaca o papel da construção de sistemas de detecção de falhas diretamente a partir de dados históricos de processos (Ge et al., 2013).

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo do estudo é a construção de um sistema combinado de detecção de falhas para um processo químico de operações contínuas. Esse sistema de detecção será baseado na combinação (*ensemble learning*) da técnica estatística multivariada denominada Análise por Componentes Principais e algumas de suas variantes. Essa combinação é baseada na estratégia de inferência Bayesiana. O objetivo final é obter uma maior capacidade de detecção de falhas, ao se comparar cada sistema em separado com aquele combinado. A motivação para essa abordagem é que nenhum sistema em particular é capaz de descrever, de modo satisfatório, as diversas características dos processos químicos, como por exemplo, não linearidade e correlação temporal (Ge e Song, 2013). O estudo de caso será o *benchmark* Tennessee (Dows e Vogel, 1993), ainda o mais utilizado pela comunidade de Engenharia Química em todo o mundo.

Para o sistema combinado de detecção de falhas, este trabalho considerou as técnicas PCA, KPCA e DPCA. A Análise por Componentes principais (PCA) é a técnica mais usual em monitoramento de processos químicos, com potencial de lidar com processos multivariados, altamente correlacionados, e com ruídos. A sua limitação é a consideração de linearidade, que é uma característica não usual em processos químicos (Dong e McAvoy, 1996). De modo a compensar a linearidade, a Análise por Componentes Principais via *Kernel* (KPCA; *Kernel* PCA) utiliza funções de *kernel* não lineares, o que possibilita a introdução de dependências não lineares entre variáveis. Nesse caso, tem-se um mapeamento não linear que projeta os dados de entrada em um espaço de características de maior dimensão, onde PCA é então aplicado. Outra questão essencial para a descrição de processos químicos é a dinâmica das operações, o que também é uma limitação de PCA. Nessa direção, tem-se outra extensão de PCA, denominada Análise por Componentes Principais, Dinâmica (DPCA;

DynamicPCA), que propõe a descrição da dinâmica do processo a partir da consideração das variáveis em instantes temporais passados. Desse modo, tendo-se a técnica PCA como ponto de partida, o sistema combinado de detecção de falhas busca considerar as relações não lineares entre variáveis de processo, ao considerar a técnica KPCA, e a dinâmica do processo, ao incorporar a técnica DPCA. Para mencionar alguns estudos com o uso dessas técnicas em separado, com o objetivo de detecção de falhas em processos químicos, Wang et al. (2018) e Du e Du (2018) são exemplos de aplicações de PCA; Kallas et al. (2017), Nguyen e Golinval (2010) e Cui et al. (2008), de KPCA; e Huang e Yan (2015), Russell et al. (2000) e Ku e Storer (1995), de DPCA.

2.2. Objetivos Específicos

Listam-se, a seguir, os objetivos específicos.

- Construção de sistemas independentes de detecção de falhas baseados em PCA, KPCA e DPCA.
- Validação de cada sistema anterior de detecção de falhas a partir da Literatura disponível.
- Construção de um sistema combinado de detecção de falhas com a combinação dos sistemas anteriores a partir de uma estratégia baseada em inferência Bayesiana.
- Comparação quantitativa da capacidade de detecção de falhas entre cada sistema em separado com o sistema combinado de detecção de falhas.

3. REVISÃO BIBLIOGRÁFICA

3.1. Histórico sobre Controle de Processos

Descreve-se, inicialmente, de modo resumido, o histórico da área de controle de qualidade (INMETRO, 2017; ASQ, 2017).

O Controle de Processos é geralmente usado como sinônimo de Controle Estatístico da Qualidade, criado por Shewhart nos anos 30 do século XX. Porém, o controle de processos não é uma técnica criada pelos americanos e aperfeiçoada pelos japoneses; porém, mais antiga e abrangente. No início, o controle de processos quase não usava conceitos estatísticos, com exceção de poucos dados descritivos.

O controle consistia, essencialmente, em métodos uniformes e normas de procedimentos. As pirâmides do Egito Antigo, a construção civil da Grécia Antiga, a organização militar dos persas e a construção naval de Veneza no século XVI, são exemplos de aplicações do antigo controle de processos, que não eram menos surpreendentes que os da revolução japonesa do século XX. Essa abordagem simples foi a precursora da Administração Científica, cujos princípios foram apresentados por Frederick W. Taylor, em 1911, em sua obra, Princípios da Administração Científica. Taylor defendia o estudo científico do trabalho, a seleção e alocação de pessoal segundo aptidões individuais e requisitos das tarefas, a colaboração entre empresas e empregados, a eficiência, e a racionalização.

Com a Revolução Industrial nos séculos XVIII e XIX, artesãos passaram a executar tarefas especializadas, e donos de lojas, a supervisionar a produção. Trabalhadores adicionais asseguravam a qualidade dos produtos por meio de inspeções e auditorias, onde os produtos defeituosos eram reprocessados ou descartados. Com o ingresso dos Estados Unidos na Segunda Guerra Mundial, em 1941, a quantidade de equipamentos militares a serem produzidos demandou atenção especial à qualidade, uma vez que o desempenho dos produtos de diferentes fabricantes deveria ser compatível. O uso de técnicas de amostragem para acelerar a inspeção e a ministração de cursos de

capacitação em técnicas de controle estatístico de processos para fornecedores de armamentos são exemplos de medidas das forças armadas para aumentara qualidade dos produtos.

Após a guerra, o Japão iniciou uma revolução na qualidade de seus produtos por meio de um novo sistema de gerenciamento, com o apoio de empresas e especialistas estrangeiros, como William E. Deming e Joseph M. Juran. Esse sistema focava na melhoria dos processos organizacionais por meio dos profissionais envolvidos. Isto tornou o Japão em um grande exportador mundial de produtos de alta qualidade e baixos preços. A resposta dos Estados Unidos, que perdera grande parte dos setores do mercado internacional, veio com um sistema semelhante ao japonês, denominado Gestão da Qualidade Total (TQM; *Total Quality Management*), com a abrangência de toda a organização.

O Controle Estatístico de Processos é uma ferramenta básica em qualquer sistema atual de gestão de qualidade total. Outras iniciativas continuaram a seguir esta tendência de busca pela qualidade, como por exemplo, a ISO (*International Organization for Standardization*) série 9000.

O próximo grande acontecimento da indústria surge com o conceito de Indústria 4.0, baseado em tecnologias disruptivas. Usado pela primeira vez na Feira de Hannover, em 2011, a apresentação de seu relatório elencou uma série de recomendações para o governo alemão, com o objetivo de se planejar a implementação de redes inteligentes na indústria. Como consequência, as fábricas inteligentes teriam a capacidade e autonomia para controlar a produção de forma autônoma, de modo a facilitar, entre outras ações, manutenções, previsões de falhas e adaptações frente às mudanças não programadas de produção. O ponto de partida para a tomada dessas ações é justamente o (antigo) controle de processos (INMETRO, 2017; SAQ, 2017).

3.2. Modelos Fenomenológicos x Modelos baseados em Dados

Os modelos fenomenológicos são baseados em uma descrição puramente matemática dos fenômenos físico-químicos. Com isso, o seu

usorestringe-se, de modo geral, a processos com um número reduzido de variáveis e estados (Venkatasubramanian et al., 2003a).

Um desafio atual é a crescente complexidade dos processos químicos, que são multivariáveis, apresentam relações não-lineares, contêm ruídos de processo e de medição, e são de conhecimento apenas parcial. Todos esses fatores dificultam o uso de modelos fenomenológicos. Aliado à crescente disponibilidade de dados sobre as operações, pelo menos em indústrias químicas de médio e grande portes, o uso de modelos baseados em dados é cada vez maior. Nesse caso, de modo geral, é suficiente ter apenas um conjunto de dados, que seja representativo das condições operacionais do processo de interesse, e um conhecimento parcial a respeito, para a obtenção de uma descrição satisfatória (Venkatasubramanian et al., 2003c). Esses são os motivos para a construção de um sistema de detecção a partir de uma abordagem baseada em dados neste trabalho.

A aquisição e o armazenamento de grandes massas de dados nas indústrias em geral devem-se ao avanço das áreas de informática, banco de dados e instrumentação, com um número crescente de sensores em toda a planta. O desafio atual é justamente transformar essas quantidades massivas de dados, denominadas de *big data*, em informação relevante para as organizações, o que é reforçado pelo contexto atual de Indústria 4.0 (Amaral, 2017; Kahneman, Lovallo e Sibony, 2011). Um exemplo dessa transformação é a construção de sistemas automáticos de detecção de falhas, que é o foco deste trabalho, e uma questão ainda em aberto no meio industrial.

3.3. Monitoramento de Processos

O monitoramento refere-se à observação e ao registro regular das operações de um processo. O acúmulo de observações, devidamente estruturado ao longo do tempo, transforma-se em informação relevante para o analista. O desempenho dos processos em geral apresentou um avanço considerável com o desenvolvimento da área de monitoramento de processos. O uso de uma grande quantidade de sensores automáticos em linha, com uma

alta frequência de amostragem, contribui para aumentar o volume de informações sobre as operações à disposição para operadores e engenheiros. Como consequência, atingiram-se operações mais seguras, limpas e econômicas. Com o objetivo de se identificar desvios, essas informações são constantemente comparadas com valores de referência previamente definidos, o ponto nominal do processo. Em resumo, a melhoria da eficiência do processo, que também depende da atividade de monitoramento de processos, é uma busca contínua que visa reduzir as perdas com oscilações das operações, falhas e paradas não programadas (Qin, 2003).

Os principais fatores que causam a perda de desempenho em processos são, falhas em equipamentos, falhas na própria planta (falha humana), falhas em dispositivos de medição (sensores), e falhas em sistemas de comunicação (por exemplo, não transmissão de dados medidos) (ASM, 2005). Essa perda de desempenho é geralmente evidenciada por medições incorretas em sensores, correlações não usuais entre variáveis e oscilações durante períodos transientes no processo.

Um sistema de monitoramento é composto, basicamente, por três etapas. A primeira é a de detecção, responsável pela identificação de falhas (o foco deste trabalho); a segunda é a de diagnóstico, responsável pela identificação de sua causa; e a terceira é a de recuperação do processo, responsável por retornar o processo à operação normal ou, pelo menos, a uma condição segura (Chiang et al., 2001). As falhas podem ser abruptas, cuja manifestação é próxima a um sinal degrau, e incipientes, cujo desenvolvimento é progressivo, a partir de um sinal de baixa magnitude. O segundo tipo é o mais usual em fábrica, e mais difícil de detectar. O *benchmark* Tennessee, que é o estudo de caso usado neste trabalho, contempla ambos os tipos de falha.

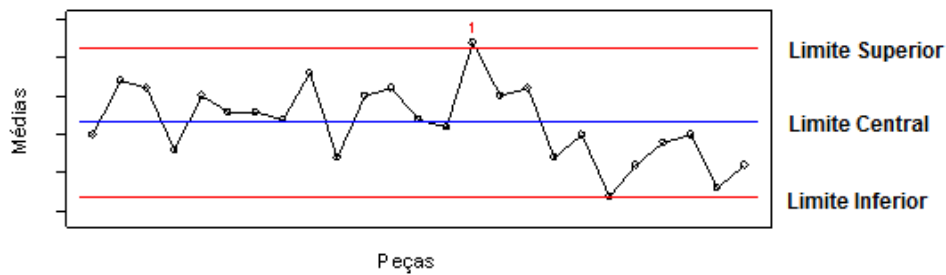
A construção de um modelo baseado em dados é geralmente a partir de uma técnica da área de estatística univariada ou multivariada (o foco deste trabalho), inteligência computacional ou processamento de sinais. A técnica estatística univariada mais usual em aplicações de monitoramento de processos químicos é a carta de controle de Shewhart. Essa área recebe a denominação de Controle Estatístico de Processos (CEP, ou SPC, *Statistical Process Control*). A técnica estatística multivariada mais comum é a Análise

por Componentes Principais (PCA), cuja área recebe a denominação de Controle Estatístico Multivariado de Processos (MSPC, *Multivariate-SPC*) (Montgomery, 2008; Wise, 1991). Após uma breve introdução sobre carta de controle, descrevem-se, a seguir, ambos os tipos de carta de controle.

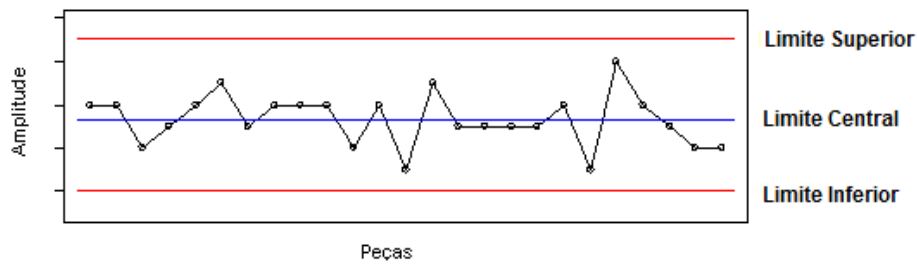
3.4. Carta de Controle

A carta de controle é um tipo de gráfico usado para o acompanhamento de um processo. Ele é geralmente composto por uma linha média (o ponto nominal do processo), uma linha inferior (Limite Inferior de Controle; LIC), e uma linha superior (Limite Superior de Controle; LSC). Uma forma usual envolve registros cronológicos regulares, com amostragens, por exemplo, por hora ou dia, de uma ou mais características, como por exemplo, média, amplitude ou proporção. O objetivo é verificar, a partir dessa característica, se o processo está sob controle estatístico, ou seja, isento de causas especiais (falhas). Em resumo, as funções desse gráfico são, mostrar evidências de uma operação sob ou fora de controle estatístico, ao indicar a ausência ou presença de causas especiais de variação (no segundo caso, tem-se a aplicação de medidas corretivas), e apresentar informações para a tomada de decisões gerenciais sobre ações de melhoria de processos (Montgomery, 2008).

Os gráficos de controle fornecem assim uma regra de decisão muito simples: pontos além dos limites de controle indicam que o processo está “fora de controle estatístico”. Caso contrário, se os pontos estão entre os limites de controle, e em uma disposição aleatória, o processo encontra-se “sob controle estatístico” (Montgomery, 2008). A Figura 1 mostra exemplos de atividades de monitoramento via gráficos de controle. Em (a), uma questão é que as observações estão entre os limites de controle, exceto por uma observação. Outra questão são os oito pontos consecutivos abaixo da linha central, que é um indicativo de falta de aleatoriedade. Em (b), a sugestão é de um processo sob controle estatístico, com as observações entre os limites de controle e com um comportamento aparentemente aleatório (Control, 1986).



(a)



(b)

Figura 1: Carta de controle (a) fora e (b) sob controle estatístico (Control, 1986).

3.5. Controle Estatístico de Processos (CEP), Univariado

O Controle Estatístico de Processos (CEP) é uma ferramenta com a finalidade de desenvolver e aplicar gráficos de controle univariados como parte de uma estratégia para a melhoria da qualidade de processos, produtos e serviços, a partir da redução de variabilidade devido à ocorrência de causas especiais.

A melhoria bem-sucedida é uma atividade orientada por gerenciamento, medindo o progresso e o sucesso em termos quantitativos. Nessa direção, a implantação de um CEP é capaz de proporcionar um retorno significativo às empresas, desde que integrado à política da organização. Os ganhos em potencial são, a informação contínua sobre o comportamento do processo, a indicação de ações para a correção de causas especiais no processo, e a informação para a melhoria contínua do processo (Montgomery, 2008).

O controle da média do processo é usualmente feito através do gráfico de controle para a média, denominado gráfico \bar{x} (usado neste trabalho). Como informação complementar, a variabilidade do processo é geralmente

monitorada por meio do gráfico de controle para o desvio-padrão, denominado gráfico s , ou do gráfico para a amplitude, denominado gráfico R . Com o primeiro gráfico, monitora-se a variabilidade dentro da amostra, e com o segundo, monitora-se a variabilidade entre amostras. O primeiro é mais usual quando o número de amostras é relativamente grande. A média μ e o desvio padrão σ são usados para construir a carta de controle para o gráfico \bar{x} , conforme a Equação 1, em que LIC é o Limite Inferior de Controle, LC, a Linha Central, e LSC, o Limite Superior de Controle. O número-multiplicador 3 corresponde a uma taxa de alarmes falsos de 0,27%, dada uma distribuição normal para os dados (Montgomery, 2008).

$$LSC = \mu + 3 \cdot \sigma \quad (1a)$$

$$LC = \mu \quad (1b)$$

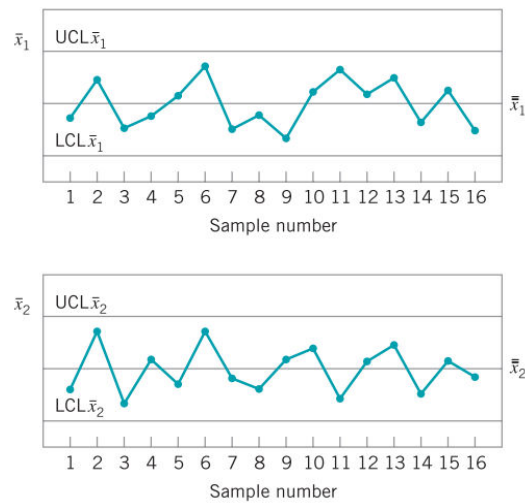
$$LIC = \mu - 3 \cdot \sigma \quad (1c)$$

No CEP univariado, cada gráfico de controle é responsável pelo monitoramento independente de uma única variável, ou seja, essa abordagem não considera a correlação espacial entre um conjunto de variáveis sob monitoramento simultâneo. Segundo Montgomery (2008), quando há múltiplos parâmetros de processo correlacionados entre si, como é o caso de processos químicos, a abordagem univariada não é adequada. A solução, nesse caso, é o uso de cartas de controle, multivariadas.

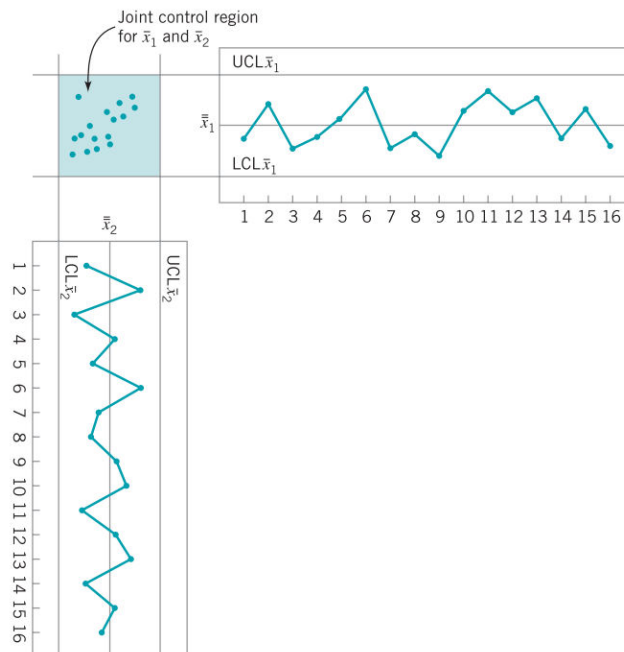
3.6. Controle Estatístico de Processos (CEP), Multivariado

As plantas das indústrias químicas modernas são equipadas com uma gama de sensores, que medem, de modo contínuo, uma diversidade de variáveis de processo. Uma característica inerente às variáveis de processo é a correlação espacial significativa, positiva ou negativa, entre pelo menos algumas delas. A Figura 2 mostra um exemplo. Ao se considerar as cartas de controle \bar{x} individuais para as variáveis X_1 e X_2 , verifica-se que ambas estão sob

controle estatístico (Figura 2(a)). Porém, ao se considerá-las em conjunto, com a definição de uma região conjunta de controle definida em função da correlação espacial entre ambas, observa-se um indicativo de uma condição fora de controle (ponto afastado na nuvem de pontos) (Figura 2(b)). Portanto, o uso de CEP univariado (subseção 3.7) não é adequado para o monitoramento de processos multivariados (Montgomery, 2008).



(a)



(b)

Figura 2: (a) Cartas de controle univariadas para as variáveis X_1 e X_2 , e (b) região de controle definida por ambas (Montgomery, 2008).

O êxito de uma aplicação de CEP multivariado depende, intrinsecamente, da capacidade do modelo, usado como o sistema de detecção de falhas, em descrever as relações entre as variáveis. O modelo mais usual em aplicações de monitoramento de processos químicos é baseado na técnica estatística multivariada denominada Análise por Componentes Principais (Yin et al., 2014). Essa técnica e algumas de suas variantes foram utilizadas neste trabalho.

4. ANÁLISE POR COMPONENTES PRINCIPAIS (PCA)

O Método das Componentes Principais foi proposto por Karl Pearson em 1901; porém, o procedimento usual foi publicado por Harold Hotelling em 1933. A Análise por Componentes Principais (PCA) é uma técnica da área de estatística multivariada cujo propósito é a redução dimensional, dada a identificação de informação redundante entre variáveis correlacionadas. Em resumo, PCA busca explicar a estrutura de variâncias e covariâncias de um conjunto de variáveis originais por meio da construção de combinações lineares dessas variáveis originais, a partir de uma transformação linear ortogonal. Essas combinações são denominadas componentes principais, não correlacionadas entre si. O objetivo final é obter k componentes, com capacidade de explicar a maior parte da variação total das p variáveis originais, em que $k \ll p$. A qualidade dessa aproximação (sinônimo de menor perda de informação) depende do número de componentes retidas, sendo usualmente mensurada por meio da proporção da variância total explicada (Jolliffe, 1986).

A descrição a seguir mostra o passo-a-passo para a obtenção do modelo PCA, o cálculo das estatísticas (métricas de monitoramento) T^2 e Q , e a definição de seus limites de controle.

Considerar a matriz $X_{[n,p]}$ (ou seja, $[X_{1[n,1]}, X_{2[n,1]}, \dots, X_{p[n,1]}]$), em que n é o número de amostras e p é o número de variáveis originais. Cada amostra é um vetor de observações (x_1, x_2, \dots, x_p) em um dado instante de tempo t . De modo geral, utilizam-se as variáveis padronizadas, ou seja, com média zero e variância um. Esse procedimento é para evitar que maiores valores de variâncias influenciem, de modo negativo, a obtenção das componentes principais. A partir das definições de média aritmética simples, de variância, que é a medida de variação de um conjunto de dados em relação à sua média, e de covariância, que é a medida de covariação entre dois conjuntos de dados (Equações 2), considerar a matriz (simétrica) de variâncias e covariâncias das variáveis originais, $\Sigma_{[p,p]}$ (Equação 3).

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2a)$$

$$S_{x_i x_i}^2 = \frac{\sum_{k=1}^n (x_{i,k} - \bar{x}_i)^2}{n-1} \quad (2b)$$

$$S_{x_i x_j}^2 = \frac{\sum_{k=1}^n (x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{n-1} \quad (2c)$$

$$\Sigma = \begin{bmatrix} S_{x_1 x_1}^2 & S_{x_1 x_2}^2 & \cdots & S_{x_1 x_p}^2 \\ S_{x_2 x_1}^2 & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & S_{x_{p-1} x_p}^2 \\ S_{x_p x_1}^2 & \cdots & S_{x_p x_{p-1}}^2 & S_{x_p x_p}^2 \end{bmatrix} \quad (3)$$

Dado a matriz quadrada $\Sigma_{[p,p]}$ e a Equação 4, com o vetor-coluna w e a constante λ , o polinômio característico de $\Sigma_{[p,p]}$ é dado pela Equação 5, que, para uma solução não-trivial (isto é, $(\Sigma - \lambda I)w = 0$, com $w \neq 0$), deve ser igual a zero, em que $I_{[p,p]}$ é a matriz identidade. Com a obtenção do vetor de autovalores ($\lambda_{[p,1]}$), obtém-se, para cada autovalor, o autovetor correspondente (w), e então, a matriz de autovetores ortonormais ($W_{[p,p]}$). Cada um dos pesos (*loads*; $w_{i,j}$, $i = 1, 2, \dots, p$ e $j = 1, 2, \dots, p$) de um autovetor (w_i) é associado a uma variável original em particular (X_j), conforme a Equação 6, segundo uma relação linear.

$$\Sigma \cdot w = \lambda \cdot w \quad (4)$$

$$\det(\Sigma - \lambda I) \quad (5)$$

$$w_i = w_{i,1} \cdot X_1 + w_{i,2} \cdot X_2 + \cdots + w_{i,p} \cdot X_p \quad (6)$$

O passo seguinte é ordenar os autovetores (as componentes principais) em ordem decrescente dos respectivos autovalores (as variâncias das componentes), e então definir o valor de k , ou seja, o número de componentes a se reter, conforme o cálculo da Equação 7, em que $VEA(\%)$ é a variância explicada acumulada em percentual. Por isso, PCA é conhecido como uma

técnica de redução de dimensionalidade, com capacidade de descrever apenas o comportamento do sistema de interesse, desconsiderando-se o ruído.

$$VEA(\%) = \sum_{i=1}^k [(\lambda_i / \sum_{j=1}^p \lambda_j) \cdot 100] \quad (7)$$

Para o cálculo dos *scores*, ou seja, os valores assumidos pelas componentes principais (*PC*) retidas, utiliza-se a Equação 8. De outro modo, os *scores* são as coordenadas das projeções dos pontos originais no novo sistema de coordenadas definido pelas componentes principais retidas. Nesse ponto, ressalta-se que PCA é uma técnica que realiza um mapeamento linear de X (o espaço original) para PC (o novo sistema de coordenadas) através de W (a matriz de *loads*) (Mingoti, 2007).

$$PC_{[n,k]} = X_{[n,p]} \cdot W_{[p,k]} \quad (8)$$

Obtido o modelo PCA, dado por $(\lambda_{[k,1]}, W_{[p,k]})$, o próximo passo é o cálculo das estatísticas (métricas de monitoramento) T^2 e Q . A primeira é calculada conforme a Equação 9, que representa o quadrado da distância de Mahalanobis de uma observação no subespaço definido pelo modelo PCA. Para o cálculo da segunda métrica, é necessário, inicialmente, calcular os valores das variáveis originais a partir de sua reconstrução ($X_{[n,p]}^*$) usando apenas as k componentes principais retidas, conforme a Equação 10. Na sequência, calculam-se os resíduos ($E_{[n,p]}$), que é dado pela diferença entre os valores originais ($X_{[n,p]}$) e aqueles reconstruídos (Equação 11). Por fim, obtém-se o valor de $(Q_{[1,1]})$, dado os resíduos $([e_1, e_2, \dots, e_p])$ de um vetor de observações (x_1, x_2, \dots, x_p) , conforme a Equação 12. Essa estatística representa a distância Euclidiana no subespaço definido pelo modelo PCA.

$$T^2 = \sum_{i=1}^k \frac{PC_i^2}{\lambda_i} \quad (9)$$

$$X_{[n,p]}^* = PC_{[n,k]} \cdot W_{[k,p]}^T \quad (10)$$

$$E_{[n,p]} = X_{[n,p]} - X_{[n,p]}^* \quad (11)$$

$$Q_{[1,1]} = \sum_{i=1}^p e_i^2 \quad (12)$$

Para o cálculo dos limites de controle (T_α^2 e Q_α) para as estatísticas T^2 e Q (ou Erro de Predição Quadrático; SPE, *Squared Prediction Error*), pode-se usar, respectivamente, as Equações 13 e 14, em que α é o nível de significância, c_α é o desvio normal padrão correspondente ao percentil $(1 - \alpha)$, $\theta_i = \sum_{j=k+1}^n \lambda_j^{2i}$, $i = 1, 2, 3$, e $h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$ (Russell et al., 2000).

$$T_\alpha^2 = \frac{k(n^2-1)}{n(n-k)} F_{k,n-k,\alpha} \quad (13)$$

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (14)$$

4.1. Análise por Componentes Principais via *Kernel* (KPCA)

A técnica PCA é baseada em um mapeamento (transformação) linear. Essa característica é um fator limitante à descrição de sistemas não lineares, como é o caso usual em processos químicos (Dong e McAvoy, 1996). Esse ponto foi a motivação para a proposta da técnica *Kernel* PCA (KPCA), que é uma extensão não linear de PCA (Scholkopf et al., 1999, 1998).

O princípio de KPCA é mapear o espaço de entrada X em um espaço de características F , de maior dimensão, através de funções de *kernel* não lineares (Figura 3); e então, aplica-se em F , o PCA linear. Os autovetores obtidos, denominados componentes principais de *kernel*, são capazes de refletir características não lineares do sistema físico. A transformação não linear é dada pelo uso da função *kernel*, que possibilita a extração de dependências não lineares entre as variáveis (Scholz et al., 2005). Têm-se

diferentes funções *kernel*, neste trabalho, adotou-se a gaussiana (Equação 15), que é a mais usual em aplicações, em que x é o vetor de observações, e x_j e $2\sigma^2$, são, respectivamente, o centro e largura (*width*) da j -ésima função *kernel*. Portanto, KPCA é uma Análise (não linear) por Componentes Principais.

$$k(x, x_j) = \exp(-\|x - x_j\|^2 / 2\sigma^2) \quad (15)$$

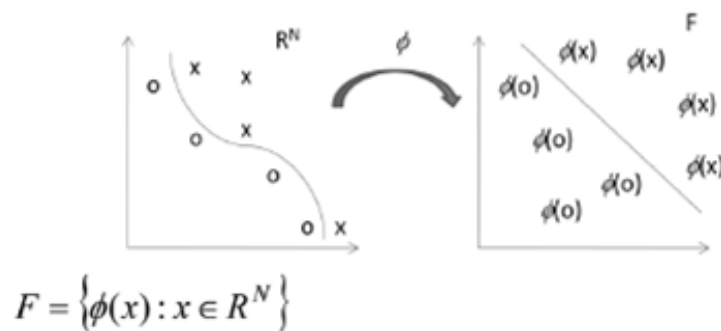


Figura 3: Mapeamento entre o espaço de entrada (R^N) e o espaço de características (F) através de uma função de *kernel* (ϕ), em que x é o vetor de observações originais (Scholkopf et al., 1999).

Após a obtenção das componentes principais, conforme descrição anterior para PCA, define-se o número de componentes (k) a se reter, e calculam-se as estatísticas T^2 e Q (Cho et al., 2005; Lee et al., 2004).

4.2. Análise por Componentes Principais, Dinâmica (DPCA)

O ponto de partida em PCA é a análise da estrutura de correlações da matriz de variáveis ($X_{[n,p]}$), em que n é o número de observações e p é o número de variáveis. Na ausência de auto-correlação (dado uma variável), a estrutura resultante contém as informações relevantes sobre o processo para uma análise satisfatória via PCA. Caso contrário, é necessário completar essa estrutura com a informação temporal (Russell et al., 2000).

Essa foi a motivação para a proposição da Análise por Componentes Principais, Dinâmica (DPCA; *Dynamic PCA*), que é uma extensão da Análise por Componentes Principais (PCA). A consideração da dinâmica do processo é um fator importante para aumentar a chance de detecção de falha, e ainda, para o alcance de uma detecção precoce. Para tal, altera-se a estrutura de correlações com a consideração da dependência temporal para as variáveis (auto-correlação), com o uso não apenas dos valores em t (um instante temporal), mas também em $(t - 1)$, $(t - 2)$, e assim por diante. Com isso, um parâmetro a se definir é o tamanho da janela temporal, ou seja, o número de atrasos (*lags*) nas variáveis. Em resumo, passa-se a se analisar uma estrutura de padrões temporais de comportamento. A seguir, a matriz de dados a se considerar como a informação de entrada para uma análise via DPCA, em que X_t^T é o vetor transposto de dimensão p , no instante t , e l é o número de atrasos (*lags*) nas variáveis originais. A diferença em relação ao modelo PCA é apenas sobre a matriz de dados ($X_{[n,p]}$), sendo a sua implementação igual àquela para PCA (Russell et al., 2000).

$$X(l) = \begin{bmatrix} X_t^T & X_{t-1}^T & \dots & X_{t-l}^T \\ X_{t-1}^T & X_{t-2}^T & \dots & X_{t-l-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ X_{t-n+l}^T & X_{t-n+l-1}^T & \dots & X_{t-n}^T \end{bmatrix}$$

4.3. Combinação entre Técnicas

Ensemble learning é o processo de combinar os diferentes resultados obtidos a partir de uma série de modelos, identificados de modo independente. O objetivo é compensar a limitação de cada modelo, e assim, aumentar a performance do sistema, seja de regressão, classificação (o caso deste trabalho), ou agrupamento, em relação aos modelos em separado. Para se alcançar esse propósito de modo satisfatório, é aconselhável usar modelos com capacidades diferentes entre si (Yu et al., 2016; Kuncheva et al., 2003). Essa diferença pode ser a partir de variações no conjunto de dados de treinamento ou a partir dos próprios modelos (Li e Yang, 2015). Neste trabalho,

consideraram-se diferentes modelos, a saber, PCA, KPCA e DPCA, com as respectivas capacidades de transformações linear e não-linear, e de auto-correlação.

Uma maneira de se combinar os resultados é a partir de uma estratégia de votação; por exemplo, ao se estabelecer um número mínimo de modelos com indicação de falha, para então se considerar a sua presença. Neste trabalho, empregou-se a estratégia de *ensemble learning* baseada em inferência Bayesiana, conforme Li e Yang (2015). Esses autores combinaram uma série de modelos KPCA, cada um com um valor distinto para o parâmetro largura das funções gaussianas. No presente trabalho, a combinação é baseada em modelos diferentes: PCA, KPCA e DPCA. Antes de se apresentar a estratégia de *ensemble learning* baseada em inferência Bayesiana, descreve-se, de modo resumido, o seu ponto de partida, o Teorema de Bayes.

4.3.1. Teorema de Bayes

Segue a motivação para o Teorema de Bayes. A probabilidade de produção de peças não-conformes (2º evento) pela máquina A (1º evento) é um cálculo direto, isto é, $P(\text{peças não – conformes} | \text{máquina A})$. Porém, dada a ocorrência e então o conhecimento de um evento (por exemplo, produção de uma peça não-conforme), deseja-se considerar essa informação para se calcular a probabilidade de ter sido produzida pela máquina A, entre um conjunto de máquinas (por exemplo, A, B e C), ou seja, $P(\text{máquina A} | \text{peça não – conforme})$. Para responder a essa questão, Thomas Bayes (1701-1761) formulou o denominado Teorema de Bayes. Conforme a descrição anterior, a informação é apresentada a partir de probabilidades condicionais. A probabilidade condicional fornece a probabilidade de um evento A (por exemplo, uma falha), dado a ocorrência de um evento B (ou condição, ou evidência), ou seja, $P(A|B)$. Da definição de probabilidade condicional, e pelo teorema da multiplicação, pode-se escrever, conforme a Equação 16.

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B \cap A) = P(B|A) \cdot P(A) \quad (16)$$

Ao considerar o segundo e o último termos, pode-se escrever a Equação 17, denominada Teorema de Bayes, a partir da qual é possível obter $P(A|B)$ em termos de $P(B|A)$ (Montgomery e Runger, 2009)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \text{ com } P(B) > 0 \quad (17)$$

4.3.2. Ensemble learning baseado em inferência Bayesiana

Após a identificação de cada modelo em separado e a obtenção das estatísticas T^2 e Q , conforme as subseções 4.1, 4.2 e 4.3, procede-se à combinação dos resultados, conforme a sequência de cálculos a seguir.

Probabilidades a priori: Definem-se, inicialmente,

- $P_{T^2}^i(N)$ e $P_Q^i(N)$: Probabilidades, *a priori*, de se ter operação normal, para as métricas de monitoramento T^2 e Q , respectivamente, dado o modelo i ($= 1$ (PCA), 2 (KPCA) e 3 (DPCA)), iguais a $(1 - \alpha)$,
- e $P_{T^2}^i(F)$ e $P_Q^i(F)$: Probabilidades, *a priori*, de se ter uma operação com falha, considerando-se as métricas de monitoramento T^2 e Q , respectivamente, dado o modelo i ($=1,2,3$), iguais a α (nível de significância, a taxa de alarmes falsos).

Termos de verossimilhança: Em seguida, definem-se para T^2 e Q , dado o modelo i , os respectivos termos de verossimilhança, considerando-se as operações normal (N) e com falha (F), em que x é o vetor de observações em um dado instante de tempo t , e LSC, o limite superior de controle.

- $P_{T^2}^i(x|N) = \exp\left(-T^{2(i)}/T_{LSC}^{2(i)}\right)$
- $P_Q^i(x|N) = \exp\left(-Q^{(i)}/Q_{LSC}^{(i)}\right)$
- $P_{T^2}^i(x|F) = \exp\left(-T_{LSC}^{2(i)}/T^{2(i)}\right)$
- $P_Q^i(x|F) = \exp\left(-Q_{LSC}^{(i)}/Q^{(i)}\right)$

Probabilidades a posteriori: A partir do Teorema de Bayes, obtêm-se as probabilidades *a posteriori* relativas a T^2 e Q , dado o modelo i , para os cenários de operação normal e de operação com falha, em que, pelo teorema da adição e teorema da probabilidade total, $P_{T^2}^i(\mathbf{x}) = P_{T^2}^i(\mathbf{x}|N) \cdot P_{T^2}^i(N) + P_{T^2}^i(\mathbf{x}|F) \cdot P_{T^2}^i(F)$ e $P_Q^i(\mathbf{x}) = P_Q^i(\mathbf{x}|N) \cdot P_Q^i(N) + P_Q^i(\mathbf{x}|F) \cdot P_Q^i(F)$, que são as probabilidades de se ter o vetor de observações \mathbf{x} em particular, relativas às estatísticas T^2 e Q . Nesse ponto, pode-se observar que os valores originais de T^2 e Q , obtidos a partir de cada modelo, foram transformados em probabilidades de falha.

- $P_{T^2}^i(N|\mathbf{x}) = \frac{P_{T^2}^i(\mathbf{x}|N) \cdot P_{T^2}^i(N)}{P_{T^2}^i(\mathbf{x})}$
- $P_Q^i(N|\mathbf{x}) = \frac{P_Q^i(\mathbf{x}|N) \cdot P_Q^i(N)}{P_Q^i(\mathbf{x})}$
- $P_{T^2}^i(F|\mathbf{x}) = \frac{P_{T^2}^i(\mathbf{x}|F) \cdot P_{T^2}^i(F)}{P_{T^2}^i(\mathbf{x})}$
- $P_Q^i(F|\mathbf{x}) = \frac{P_Q^i(\mathbf{x}|F) \cdot P_Q^i(F)}{P_Q^i(\mathbf{x})}$

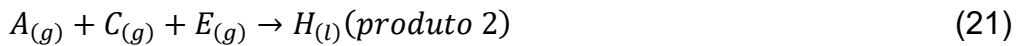
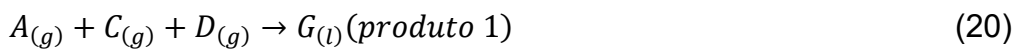
Ensemble métricas: A partir do conhecimento das probabilidades *a posteriori* (probabilidades de falha), calculam-se as combinações dos resultados a partir de *ensemble* métricas para T^2 (ET^2 ; Equação 18) e Q (EQ ; Equação 19), a partir de uma combinação ponderada, em que M é o número de modelos; neste trabalho, igual a 3 (PCA, KPCA e DPCA). Observa-se que um modelo com maior probabilidade de falha contribui, relativamente, com um maior peso. Conforme Li e Yang (2015), se $ET^2 < \alpha$ e $EQ < \alpha$, então operação normal; caso contrário, operação com falha. Neste trabalho, calculou-se um LSC para cada métrica, ET^2 e EQ , com uma taxa de alarmes falsos de 1%.

- $ET^2 = \sum_{i=1}^M \frac{P_{T^2}^{2(i)}(F|\mathbf{x})}{\sum_{j=1}^M P_{T^2}^j(F|\mathbf{x})}$ (18)

- $EQ = \sum_{i=1}^M \frac{P_Q^{2(i)}(F|\mathbf{x})}{\sum_{j=1}^M P_Q^j(F|\mathbf{x})}$ (19)

5. BENCHMARK TENNESSEE

O *benchmark* Tennessee é um estudo de caso baseado em uma indústria química real, com o objetivo de avaliação de diferentes estratégias de controle e de monitoramento de processos. Ainda hoje, é o principal *benchmark* utilizado pela comunidade de Engenharia Química em todo o mundo. O processo é descrito pelo conjunto de reações químicas conforme as Equações 20 a 23, de natureza irreversível e exotérmica. Há um total de oito componentes: quatro reagentes: A, C, D e E, um inerte: B, dois produtos: G e H, e um subproduto: F. A Figura 4 mostra o fluxograma de processo, composto por cinco equipamentos principais: o reator, o condensador de produto, o separador líquido-vapor, o compressor de reciclagem e o *stripper* de produto. Há um total de 52 variáveis de processo e de laboratório, coletadas em um intervalo de amostragem de 3 minutos (Downs e Vogel, 1993).



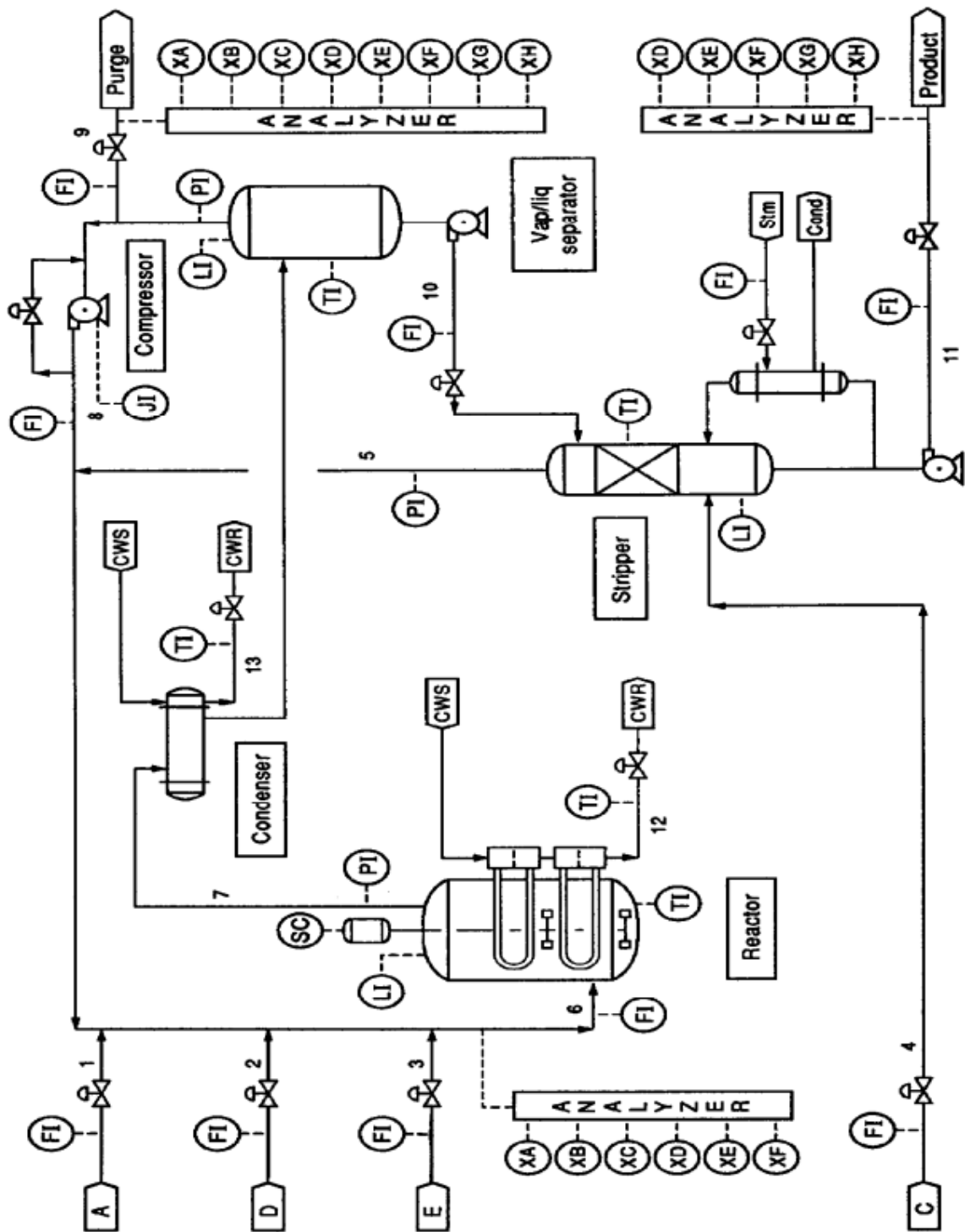


Figura 4: Fluxograma de processo do *benchmark* Tennessee (Downs e Vogel, 1993).

Há 21 cenários de falhas à disposição, conforme a Tabela 1, que contemplam diversas situações usuais em processos industriais em geral. O início de qualquer falha é em t (tempo) = 8h, correspondente à observação 160.

Falha	Descrição	Tipo
1	Razão de alimentação A/C, B composição constante	Degrau
2	Composição B, razão A/C constante	Degrau
3	Temperatura de alimentação de D(linha2)	Degrau
4	Temperatura da água de resfriamento na entrada do reator	Degrau
5	Temperatura da água na entrada do condensador	Degrau
6	Perda de alimentação A(linha1)	Degrau
7	Queda de pressão de C - diminuição de disponibilidade	Degrau
8	Composições alimentação A, B e C(linha 4)	Variação aleatória
9	Temperatura de alimentação de D(linha2)	Variação aleatória
10	Temperatura de alimentação de C(linha4)	Variação aleatória
11	Temperatura da água de resfriamento na entrada do reator	Variação aleatória
12	Temperatura da água na entrada do condensador	Variação aleatória
13	Cinética do reator	Alteração lenta
14	Válvula de água de resfriamento	Emperramento
15	Válvula de água do condensador	Emperramento
16	Desconhecido	–
17	Desconhecido	–
18	Desconhecido	–
19	Desconhecido	–
20	Desconhecido	–
21	Válvula linha 4 fixa em estado estacionário	Posição fixa

Tabela 1:Conjunto de falhas no *benchmark* Tennessee
(Russell et al., 2000; Dows e Vogel, 1993).

Descreve-se, a seguir, a metodologia referente à construção e validação de cada sistema de detecção de falhas em separado, e de construção do sistema combinado de detecção de falhas.

6. METODOLOGIA

Conforme a proposta deste trabalho, o objetivo é construir e combinar três sistemas de detecção de falhas baseados em PCA, DPCA e KPCA. Para os dois primeiros, construíram-se os respectivos *scripts* em Matlab, e para o terceiro, empregou-se a sub-rotina em disponível em *Matlab Toolbox for Dimensionality Reduction* (van der Maaten, 2018). Descrevem-se, a seguir, as etapas de construção e validação de cada sistema de detecção de falhas, e de combinação dos resultados obtidos com esses sistemas.

6.1. Etapa 1: Construção e Validação dos Sistemas de Detecção de Falhas

Apresenta-se, a seguir, a sequência de passos para a obtenção de cada sistema de sistema de detecção de falhas, a saber, PCA, KPCA e DPCA.

Passo 1: Obtenção do modelo a partir de uma matriz de variáveis originais ($X_{[n=500,p=52]}$), representativa de condições normais de operação, em que n é o número de observações e p é o número de variáveis originais, conforme definição no *benchmark* Tennessee. As variáveis são inicialmente padronizadas com média zero e variância um.

Passo 2: Definição dos limites superiores de controle para as estatísticas (métricas de monitoramento) T^2 e Q , a partir de um segundo conjunto de dados também característico de condições normais de operação ($X_{[n=960,p=52]}$), conforme definição no *benchmark* Tennessee. Para a determinação dos limites, não se empregou as Equações 13 e 14, onde se tem a suposição de distribuição normal multivariada. Neste trabalho, com a definição da taxa de alarmes falsos em 1%, conforme Russell et al. (2000), calcularam-se os valores dos limites para T^2 e Q a partir do respectivo cálculo do percentil 99.

Passo 3: Validação das cartas de controle para T^2 e Q. Com os passos 1 e 2, obtêm-se as respectivas cartas de controle para T^2 e Q. Na sequência, utilizou-se um resultado clássico, disponível na Literatura, para a validação de ambas as cartas. Para PCA e DPCA, utilizou-se o artigo de Russell et al. (2000), e para KPCA, utilizou-se o livro de Ge e Song (2013; capítulo 5).

6.2. Etapa 2: Combinação dos Sistemas de Detecção de Falhas

Apresentou-se o passo-a-passo da abordagem via *ensemble learning* baseada em inferência Bayesiana na subseção 4.3.2, a partir dos cálculos das probabilidades *a priori*, dos termos de verossimilhança, das probabilidades *a posteriori*, e por fim, das *ensemble* métricas relativas a T^2 e Q (ET^2 e EQ , respectivamente).

7. RESULTADO E DISCUSSÃO

7.1. Etapa 1: Construção e Validação dos Sistemas de Detecção de Falhas

Apresenta-se, de modo conjunto, para cada passo desta etapa, os resultados dos modelos PCA, KPCA e DPCA.

Passo 1: Obtenção do modelo a partir de uma matriz de variáveis originais $X_{[n,p]}$, características de condições normais de operação.

As Tabelas 2 a 4 apresentam, respectivamente, para os modelos PCA, KPCA (D (dimensão do espaço de características) = 21, e $k = 13$) e DPCA (l (número de *lags*) = 3), os conjuntos de autovalores, variância explicada e variância explicada acumulada.

CP	Auto valor	Variância Explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)
1	6,607	12,707	12,707	19	0,995	1,913	71,856	37	0,451	0,868	96,458
2	3,933	7,564	20,271	20	0,961	1,849	73,705	38	0,443	0,853	97,311
3	2,809	5,403	25,673	21	0,936	1,800	75,505	39	0,391	0,752	98,063
4	2,331	4,483	30,156	22	0,896	1,724	77,229	40	0,337	0,647	98,710
5	2,195	4,221	34,377	23	0,874	1,682	78,910	41	0,284	0,547	99,257
6	2,084	4,007	38,384	24	0,830	1,596	80,506	42	0,213	0,410	99,667
7	1,934	3,719	42,103	25	0,811	1,559	82,065	43	0,080	0,153	99,821
8	1,735	3,336	45,439	26	0,772	1,484	83,549	44	0,046	0,089	99,910
9	1,626	3,127	48,566	27	0,765	1,470	85,019	45	0,021	0,041	99,951
10	1,503	2,890	51,456	28	0,721	1,386	86,405	46	0,011	0,021	99,972
11	1,404	2,699	54,155	29	0,698	1,342	87,747	47	0,008	0,015	99,987
12	1,287	2,475	56,630	30	0,661	1,271	89,018	48	0,004	0,007	99,994
13	1,255	2,413	59,043	31	0,631	1,214	90,232	49	0,003	0,006	100,000
14	1,227	2,359	61,401	32	0,611	1,175	91,407	50	0,000	0,000	100,000
15	1,195	2,297	63,699	33	0,582	1,119	92,525	51	0,000	0,000	100,000
16	1,122	2,157	65,856	34	0,580	1,115	93,641	52	0,000	0,000	100,000
17	1,072	2,062	67,918	35	0,523	1,005	94,646				
18	1,053	2,025	69,943	36	0,492	0,945	95,591				

Tabela 2: Autovalores em ordem decrescente para o modelo PCA.

CP	Auto valor	Variância Explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)
1	0,253	20,291	20,291	11	0,059	4,778	90,943	21	0,000	0,004	100,000
2	0,160	12,826	33,117	12	0,048	3,825	94,769				
3	0,105	8,406	41,524	13	0,034	2,753	97,522				
4	0,099	7,974	49,498	14	0,025	2,003	99,525				
5	0,095	7,659	57,156	15	0,003	0,272	99,797				
6	0,082	6,562	63,718	16	0,002	0,170	99,966				
7	0,077	6,156	69,874	17	0,000	0,013	99,979				
8	0,072	5,754	75,628	18	0,000	0,007	99,986				
9	0,070	5,650	81,278	19	0,000	0,005	99,991				
10	0,061	4,888	86,166	20	0,000	0,004	99,996				

Tabela 3: Autovalores em ordem decrescente para o primeiro modelo KPCA, usado para fins de validação.

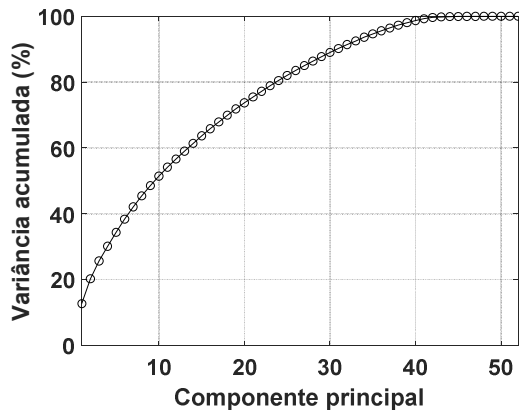
CP	Auto valor	Variância Explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)
1	25,322	12,174	12,174	71	0,802	0,385	86,950	141	0,092	0,044	99,379
2	14,665	7,050	19,224	72	0,795	0,382	87,332	142	0,089	0,043	99,421
3	7,177	3,451	22,675	73	0,764	0,367	87,699	143	0,084	0,040	99,462
4	5,315	2,555	25,230	74	0,740	0,356	88,055	144	0,081	0,039	99,501
5	5,089	2,447	27,676	75	0,726	0,349	88,404	145	0,074	0,036	99,536
6	4,790	2,303	29,979	76	0,714	0,343	88,747	146	0,071	0,034	99,570
7	4,241	2,039	32,018	77	0,703	0,338	89,085	147	0,069	0,033	99,603
8	4,018	1,932	33,950	78	0,686	0,330	89,415	148	0,066	0,032	99,635
9	3,796	1,825	35,775	79	0,678	0,326	89,741	149	0,061	0,030	99,664
10	3,728	1,792	37,567	80	0,664	0,319	90,060	150	0,056	0,027	99,691
11	3,626	1,743	39,310	81	0,649	0,312	90,373	151	0,055	0,026	99,717
12	3,299	1,586	40,896	82	0,633	0,304	90,677	152	0,050	0,024	99,741
13	3,152	1,515	42,411	83	0,612	0,294	90,971	153	0,047	0,022	99,764
14	3,030	1,457	43,868	84	0,606	0,291	91,262	154	0,044	0,021	99,785
15	2,924	1,406	45,274	85	0,598	0,287	91,550	155	0,041	0,020	99,805
16	2,832	1,362	46,635	86	0,572	0,275	91,825	156	0,038	0,018	99,823
17	2,747	1,321	47,956	87	0,561	0,270	92,095	157	0,036	0,017	99,840
18	2,664	1,281	49,237	88	0,559	0,269	92,363	158	0,034	0,016	99,856
19	2,602	1,251	50,488	89	0,549	0,264	92,627	159	0,031	0,015	99,871
20	2,522	1,213	51,700	90	0,528	0,254	92,881	160	0,028	0,013	99,885
21	2,481	1,193	52,893	91	0,520	0,250	93,131	161	0,027	0,013	99,898
22	2,399	1,153	54,047	92	0,504	0,242	93,373	162	0,023	0,011	99,909
23	2,350	1,130	55,177	93	0,499	0,240	93,613	163	0,020	0,010	99,919
24	2,298	1,105	56,281	94	0,492	0,237	93,850	164	0,019	0,009	99,928
25	2,195	1,056	57,337	95	0,481	0,231	94,081	165	0,018	0,009	99,936
26	2,190	1,053	58,390	96	0,469	0,226	94,306	166	0,016	0,008	99,944
27	2,139	1,028	59,418	97	0,450	0,216	94,523	167	0,015	0,007	99,951
28	2,071	0,996	60,414	98	0,433	0,208	94,731	168	0,013	0,006	99,957
29	2,028	0,975	61,389	99	0,427	0,205	94,936	169	0,012	0,006	99,963
30	2,009	0,966	62,354	100	0,421	0,203	95,138	170	0,012	0,006	99,969
31	1,955	0,940	63,294	101	0,401	0,193	95,331	171	0,007	0,004	99,972
32	1,905	0,916	64,210	102	0,380	0,183	95,514	172	0,006	0,003	99,975
33	1,851	0,890	65,100	103	0,365	0,176	95,690	173	0,006	0,003	99,978
34	1,843	0,886	65,986	104	0,356	0,171	95,861	174	0,005	0,003	99,981
35	1,728	0,831	66,817	105	0,344	0,165	96,026	175	0,005	0,002	99,983

Tabela 4 (parte 1): Autovalores em ordem decrescente para o modelo DPCA.

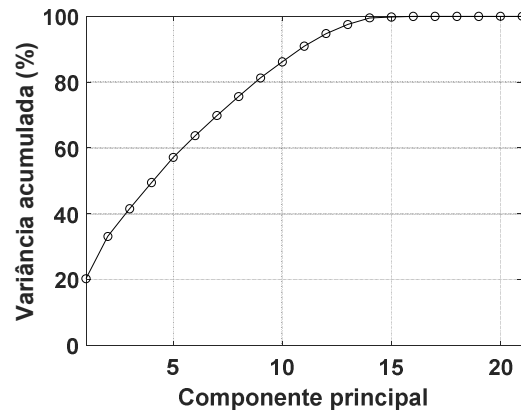
36	1,698	0,816	67,633	106	0,332	0,160	96,186	176	0,004	0,002	99,985
37	1,644	0,790	68,423	107	0,326	0,156	96,342	177	0,004	0,002	99,987
38	1,554	0,747	69,170	108	0,315	0,151	96,494	178	0,003	0,002	99,989
39	1,539	0,740	69,910	109	0,313	0,150	96,644	179	0,003	0,002	99,990
40	1,505	0,724	70,634	110	0,297	0,143	96,787	180	0,003	0,001	99,991
41	1,477	0,710	71,344	111	0,279	0,134	96,921	181	0,003	0,001	99,993
42	1,426	0,685	72,029	112	0,278	0,134	97,055	182	0,002	0,001	99,994
43	1,408	0,677	72,706	113	0,267	0,129	97,183	183	0,002	0,001	99,995
44	1,369	0,658	73,364	114	0,262	0,126	97,309	184	0,002	0,001	99,996
45	1,353	0,650	74,015	115	0,257	0,123	97,433	185	0,002	0,001	99,997
46	1,312	0,631	74,646	116	0,240	0,116	97,548	186	0,002	0,001	99,997
47	1,277	0,614	75,260	117	0,234	0,113	97,661	187	0,002	0,001	99,998
48	1,265	0,608	75,868	118	0,224	0,108	97,769	188	0,001	0,000	99,999
49	1,237	0,595	76,463	119	0,222	0,107	97,876	189	0,001	0,000	99,999
50	1,235	0,594	77,057	120	0,218	0,105	97,981	190	0,001	0,000	99,999
51	1,177	0,566	77,623	121	0,195	0,094	98,074	191	0,001	0,000	100,000
52	1,170	0,562	78,185	122	0,188	0,090	98,164	192	0,000	0,000	100,000
53	1,125	0,541	78,726	123	0,183	0,088	98,252	193	0,000	0,000	100,000
54	1,119	0,538	79,264	124	0,177	0,085	98,337	194	0,000	0,000	100,000
55	1,090	0,524	79,788	125	0,171	0,082	98,419	195	0,000	0,000	100,000
56	1,055	0,507	80,295	126	0,165	0,079	98,499	196	0,000	0,000	100,000
57	1,054	0,507	80,802	127	0,164	0,079	98,577	197	0,000	0,000	100,000
58	1,015	0,488	81,290	128	0,155	0,075	98,652	198	0,000	0,000	100,000
59	1,001	0,481	81,771	129	0,142	0,068	98,721	199	0,000	0,000	100,000
60	0,992	0,477	82,248	130	0,141	0,068	98,788	200	0,000	0,000	100,000
61	0,987	0,474	82,723	131	0,137	0,066	98,854	201	0,000	0,000	100,000
62	0,943	0,454	83,176	132	0,132	0,063	98,917	202	0,000	0,000	100,000
63	0,926	0,445	83,622	133	0,131	0,063	98,981	203	0,000	0,000	100,000
64	0,918	0,441	84,063	134	0,121	0,058	99,039	204	0,000	0,000	100,000
65	0,902	0,434	84,497	135	0,113	0,054	99,093	205	0,000	0,000	100,000
66	0,890	0,428	84,925	136	0,111	0,053	99,146	206	0,000	0,000	100,000
67	0,868	0,417	85,342	137	0,104	0,050	99,196	207	0,000	0,000	100,000
68	0,860	0,414	85,755	138	0,099	0,048	99,244	208	0,000	0,000	100,000
69	0,851	0,409	86,164	139	0,095	0,046	99,289				
70	0,832	0,400	86,564	140	0,094	0,045	99,334				

Tabela 4 (parte 2): Autovalores em ordem decrescente para o modelo DPCA.

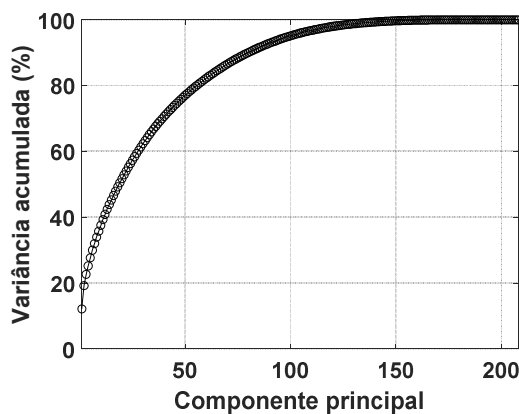
A Figura 5 mostra, de modo gráfico, a variância explicada acumulada (em %) para os modelos PCA, KPCA e DPCA.



(a) Modelo PCA.



(b) Modelo KPCA.



(c) Modelo DPCA.

Figura 5: Variância explicada acumulada (%).

Para fins de validação, o modelo PCA final é composto por 11 componentes principais (conforme definição de Russell et al. (2000)), que explicam, aproximadamente, 54% (54,155%) da variação total das variáveis originais. Para o modelo KPCA, adotou-se a função *kernel* mais usual, que é a gaussiana (Equação 24). Em relação ao parâmetro de largura ($2\sigma^2$), empregou-se o valor sugerido por Cho et al. (2005), igual a $5p$, em que p é o número de variáveis originais. O modelo KPCA final, com apenas 16 do total de 52 variáveis originais (conforme definição em Ge e Song (2013; capítulo 5)), utiliza um espaço de características de dimensão igual a 21 e um número de

componentes principais retidas (k) igual a 13, cuja capacidade de explicação da variância total é de aproximadamente 98% (97,522%). O modelo DPCA final, com um *lag* de três unidades de tempo, é composto por 29 componentes principais (conforme definição de Russell et al. (2000)), com capacidade de explicação de aproximadamente 61% (61,389%) da variação total dos originais.

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (24)$$

Passo 2: Definição dos limites superiores de controle para as estatísticas (métricas de monitoramento) T^2 e Q , a partir de um segundo conjunto de dados característico de condições normais de operação $X_{[n=960,p=52]}$.

A Tabela 5 mostra os limites de controle para os modelos PCA, KPCA e DPCA, equivalentes ao percentil 99, ou seja, dada a taxa de alarmes falsos de 1% (conforme Russell et al., 2000).

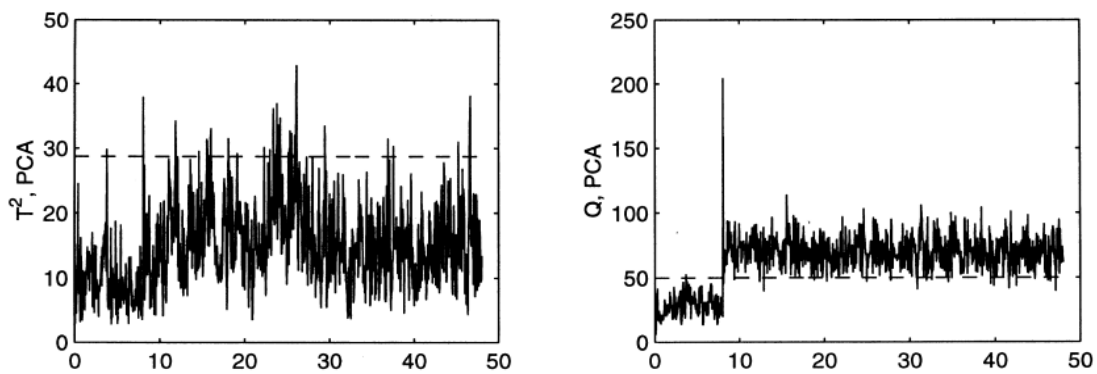
Técnica	T^2	Q
PCA	29,699	50,869
KPCA	0,070835	0,00042413
DPCA	52,354	157,630

Tabela 5: Limites superior de controle para T^2 e Q .

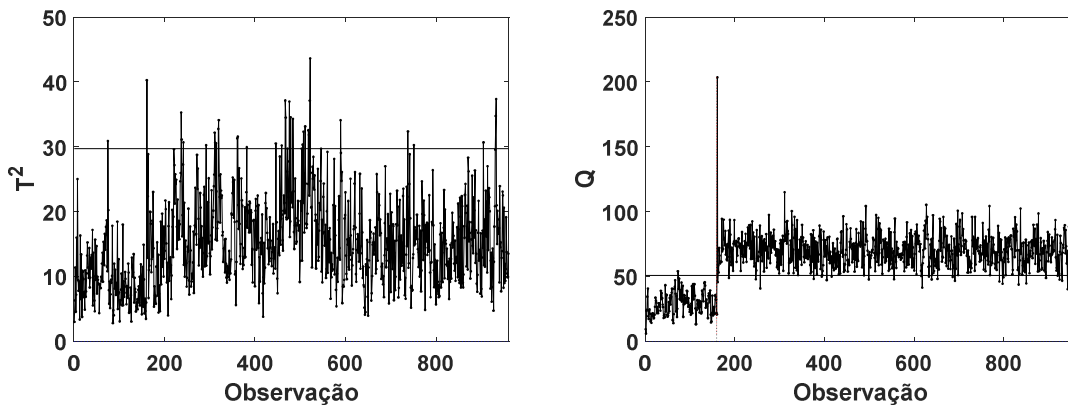
Passo 3: Validação das cartas de controle para T^2 e Q . Na sequência, utilizou-se um resultado disponível na Literatura para a validação de ambas as cartas.

Apresenta-se, inicialmente, uma validação qualitativa, e na sequência, uma validação quantitativa. Para tal, compararam-se os resultados dos modelos PCA e DPCA com aqueles em Russell et al. (2000), e do modelo KPCA, com aquele em Ge e Song (2013).

As Figuras 6 e 7, mostram, respectivamente, para as abordagens via PCA (com $k = 11$) e DPCA (com $k = 29$, e $l = 3$), os gráficos de T^2 e Q obtidos pelos respectivos modelos deste trabalho e aqueles disponíveis em Russell et al. (2000) para as falhas 4 e 5. Pode-se observar, de modo visual, a similaridade entre os pares de gráficos. Ge e Song (2013; capítulo 5) não apresentam gráficos de modo a se ter uma validação qualitativa para o modelo KPCA; porém, apresentam uma tabela de resultados, o que possibilita a sua validação quantitativa.

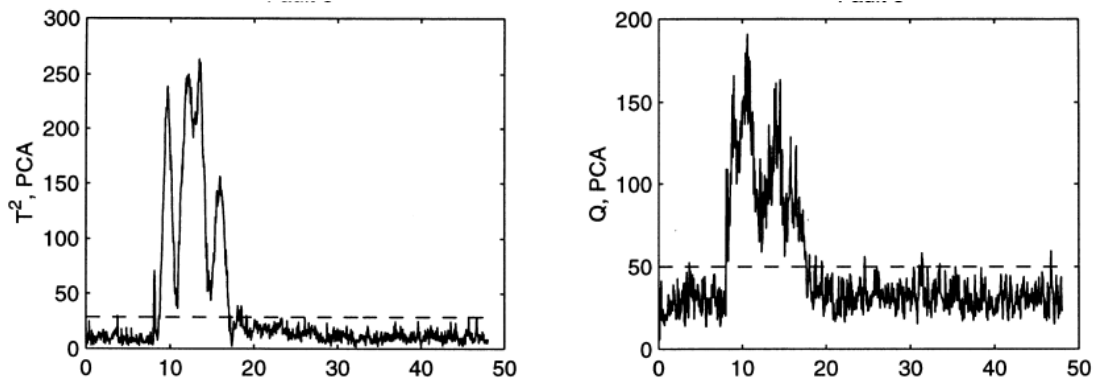


(a) Russell et al. (2000) – Falha 4.

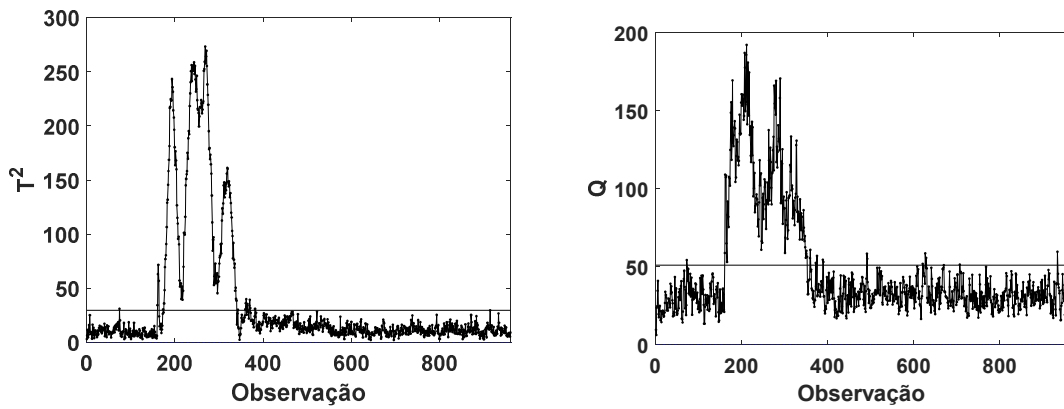


(b) Modelo PCA deste trabalho – Falha 4.

Figura 6 (parte 1): Valores T^2 e Q para as falhas 4 e 5, dado o modelo PCA.

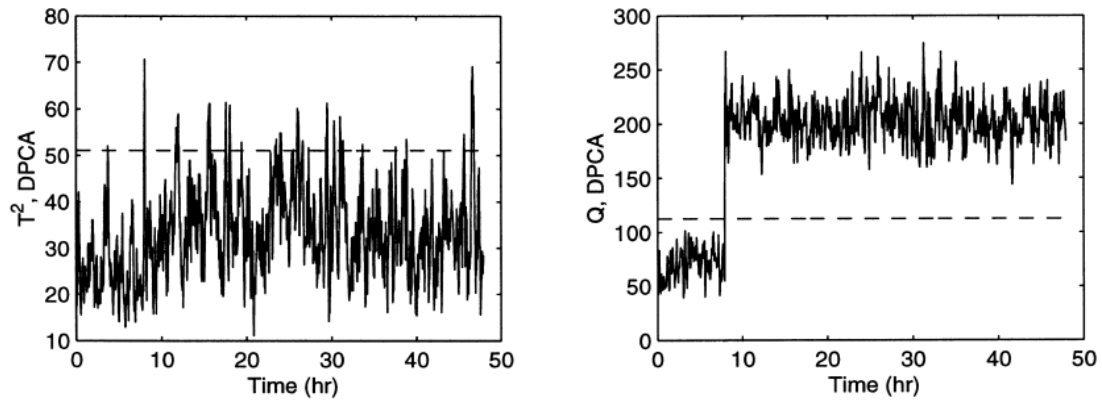


(c) Russell et al. (2000) – Falha 5.

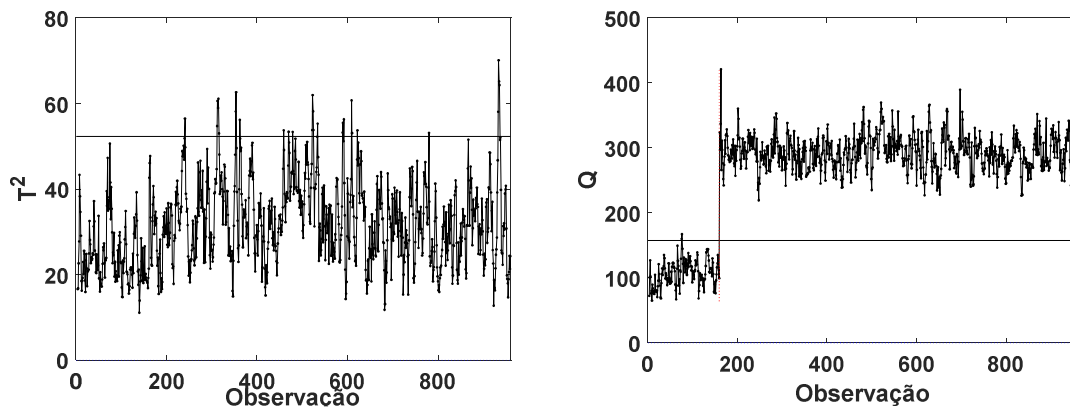


(d) Modelo PCA deste trabalho – Falha 5.

Figura 6 (parte 2): Valores T^2 e Q para as falhas 4 e 5, dado o modelo PCA.

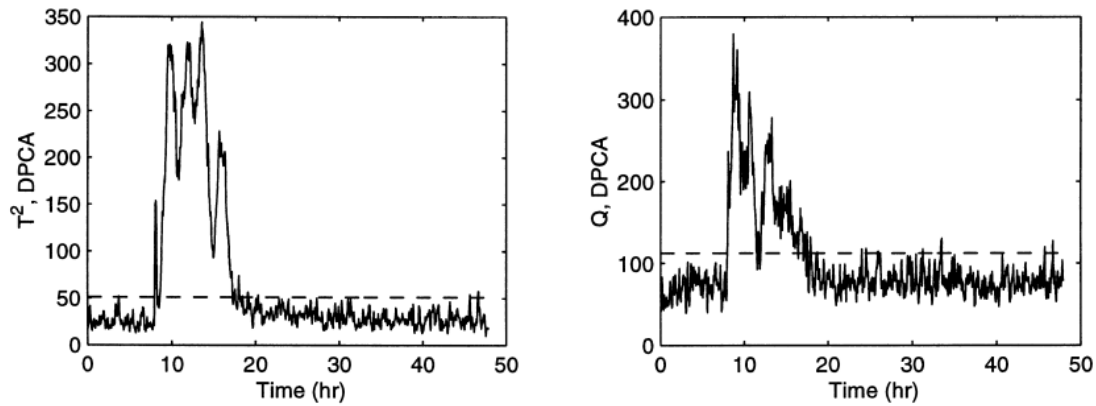


(a) Russell et al. (2000) – Falha 4.

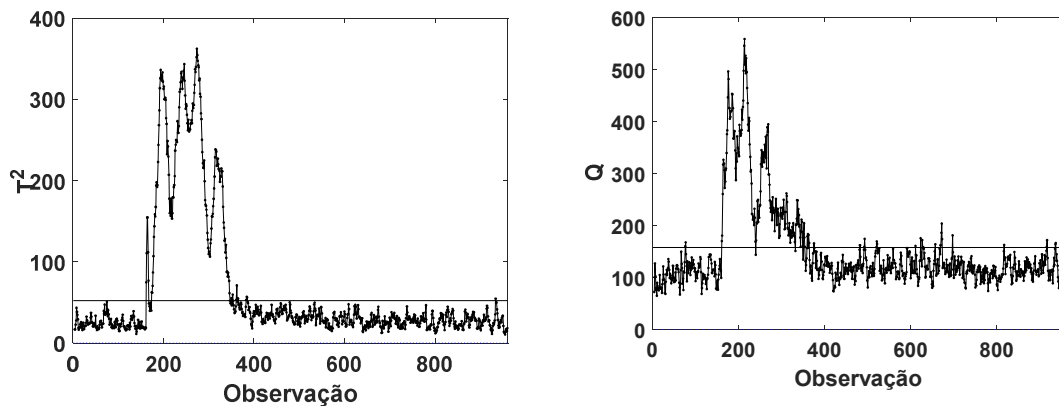


(b) Modelo DPCA deste trabalho – Falha 4.

Figura 7 (parte 1): Valores T^2 e Q para as falhas 4 e 5, dado o modelo DPCA.



(c) Russell et al. (2000) – Falha 5.



(d) Modelo DPCA deste trabalho – Falha 5.

Figura 7 (parte 2): Valores T^2 e Q para as falhas 4 e 5, dado o modelo DPCA.

As Tabelas 6 a 8 apresentam as comparações entre os resultados dos modelos PCA e DPCA deste trabalho, com aqueles obtidos em Russell et al. (2000), e do modelo KPCA, com aqueles obtidos em Ge e Song (2013), para a métrica de monitoramento denominada taxa de detecções perdidas (MDR, *Missed Detection Rate*). Essa métrica refere-se à proporção de pontos na região de operação normal após a ocorrência da falha (erro tipo II).

Em relação ao modelo PCA, a diferença absoluta (em %) é sempre inferior a 5%; na verdade, próxima de zero. Com essa análise quantitativa, em conjunto com a análise gráfica anterior, validou-se o modelo PCA construído neste trabalho. Para o modelo DPCA, a diferença absoluta (em %) é superior a 5% apenas para três valores (em *itálico*), relativos à T^2 para a falha 7 e à Q para as falhas 10 e 19. Assim, validou-se o modelo DPCA construído neste trabalho. Para o modelo KPCA, tem-se apenas quatro valores de diferença

absoluta (em %) acima de 5% (em *itálico*), relativos à estatística Q para as falhas 7, 11, 16 e 19, sendo a maioria com proximidade significativa à zero 0%. Desse modo, também validou-se o modelo KPCA obtido neste trabalho.

Falha	Modelo PCA deste Trabalho		Russell et al. (2000)		Diferença absoluta (%)	
	T ²	Q	T ²	Q	T ²	Q
1	0,009	0,004	0,008	0,003	0,074	0,075
2	0,021	0,015	0,02	0,014	0,122	0,098
3	1,000	0,990	0,998	0,991	0,200	0,099
4	0,955	0,035	0,956	0,038	0,094	0,304
5	0,777	0,743	0,775	0,746	0,153	0,318
6	0,012	0,001	0,011	0	0,148	0,125
7	0,090	0,001	0,085	0	0,489	0,125
8	0,035	0,025	0,034	0,024	0,096	0,097
9	0,994	0,981	0,994	0,981	0,024	0,027
10	0,660	0,634	0,666	0,659	0,558	2,479
11	0,790	0,347	0,794	0,356	0,374	0,893
12	0,030	0,025	0,029	0,025	0,096	0,003
13	0,061	0,046	0,06	0,045	0,117	0,119
14	0,162	0,001	0,158	0	0,430	0,125
15	0,985	0,965	0,988	0,973	0,298	0,796
16	0,828	0,737	0,834	0,755	0,628	1,842
17	0,258	0,106	0,259	0,108	0,057	0,188
18	0,114	0,102	0,113	0,101	0,061	0,137
19	0,998	0,864	0,996	0,873	0,150	0,908
20	0,700	0,543	0,701	0,55	0,063	0,693
21	0,739	0,564	0,736	0,57	0,308	0,571

Tabela 6: Modelo PCA: Validação via comparação com Russell et al. (2000) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, *Missed Detection Rate*).

Falha	Modelo DPCA deste Trabalho		Russell et al. (2000)		Diferença absoluta (%)	
	T ²	Q	T ²	Q	T ²	Q
1	0,006	0,006	0,006	0,005	0,024	0,124
2	0,020	0,012	0,019	0,015	0,098	0,252
3	0,991	0,985	0,991	0,990	0,026	0,498
4	0,965	0,001	0,939	0,000	2,604	0,125
5	0,764	0,728	0,758	0,748	0,604	2,016
6	0,014	0,001	0,013	0,000	0,073	0,125
7	0,427	0,001	0,159	0,000	26,797	0,125
8	0,029	0,026	0,028	0,025	0,071	0,122
9	0,998	0,991	0,995	0,994	0,250	0,274
10	0,584	0,609	0,580	0,665	0,427	5,576
11	0,818	0,167	0,801	0,193	1,673	2,571
12	0,011	0,016	0,010	0,024	0,124	0,777
13	0,054	0,049	0,049	0,049	0,468	0,031
14	0,055	0,002	0,061	0,000	0,607	0,250
15	0,966	0,979	0,964	0,976	0,229	0,278
16	0,800	0,668	0,783	0,708	1,725	4,008
17	0,238	0,044	0,240	0,053	0,155	0,930
18	0,114	0,101	0,111	0,100	0,261	0,112
19	0,996	0,662	0,993	0,735	0,325	7,333
20	0,635	0,464	0,644	0,490	0,854	2,558
21	0,644	0,556	0,644	0,558	0,019	0,244

Tabela 7: Modelo DPCA: Validação via comparação com Russell et al. (2000) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, *Missed Detection Rate*).

Falha	Metodo KPCA deste Trabalho		Ge e Song (2013)		Diferença absoluta (%)	
	T ²	Q	T ²	Q	T ²	Q
1	0,003	0,004	0,008	0,003	0,450	0,150
2	0,018	0,018	0,018	0,019	0,050	0,075
3	0,971	0,979	0,973	0,988	0,150	0,850
4	0,990	0,988	0,993	0,988	0,250	0,050
5	0,763	0,756	0,763	0,801	0,050	4,525
6	0,000	0,000	0,000	0,008	0,000	0,750
7	0,591	0,595	0,600	0,721	0,900	12,625
8	0,023	0,023	0,030	0,058	0,700	3,450
9	0,975	0,970	0,975	0,979	0,000	0,875
10	0,608	0,338	0,619	0,379	1,075	4,075
11	0,575	0,609	0,595	0,980	2,000	37,100
12	0,010	0,013	0,011	0,060	0,125	4,700
13	0,050	0,055	0,056	0,050	0,625	0,500
14	0,000	0,001	0,000	0,001	0,000	0,025
15	0,923	0,938	0,944	0,984	2,075	4,575
16	0,703	0,686	0,721	0,913	1,825	22,650
17	0,040	0,064	0,073	0,066	3,250	0,225
18	0,099	0,103	0,103	0,100	0,350	0,300
19	0,881	0,864	0,906	0,956	2,525	9,225
20	0,693	0,490	0,716	0,498	2,325	0,750
21	0,574	0,584	0,581	0,628	0,725	4,350

Tabela 8: Método KPCA: Validação via comparação com Ge e Song (2013) através da métrica de monitoramento: Taxa de detecções perdidas (MDR, *Missed Detection Rate*).

7.2. Etapa 2: Combinação dos Sistemas de Detecção de Falhas

Após a validação dos *scripts* de construção de cada modelo em separado (subseção anterior), o passo seguinte é a obtenção de cada modelo com um percentual adequado de explicação da variância total dos dados em problemas de engenharia, neste trabalho, igual a 85%, e posteriormente, a combinação de seus resultados de modo a se alcançar um melhor desempenho em relação à detecção das falhas, conforme o objetivo proposto neste trabalho.

De modo similar ao passo 1 da Etapa 1 anterior de Construção e validação dos sistemas de detecção de falhas, definiu-se um modelo PCA com 27 componentes principais, com capacidade de explicação de 85,019% da variação total (ver Tabela 2). Para o modelo KPCA, utilizaram-se as 52 variáveis originais, adotou-se um parâmetro de largura ($2\sigma^2$) igual a $12,5p$ (Li e Yang,

2015), em que p é o número de variáveis originais, e definiu-se um espaço de características de dimensão igual a 42 e um número de componentes principais retidas (k) igual a 27, cuja capacidade de explicação é igual 85,303% da variância total (ver Tabela 9). Para o modelo DPCA, usou-se o mesmo *lag* de três unidades de tempo, com a retenção de 67 componentes principais, cuja capacidade de explicação é de 85,342% da variação total dos originais (ver Tabela 4).

CP	Auto valor	Variância Explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)	CP	Auto valor	Variância explicada (%)	Variância Explicada Acumulada (%)
1	0,008	12,749	12,749	15	0,001	2,305	63,911	29	0,001	1,346	88,040
2	0,005	7,589	20,338	16	0,001	2,164	66,075	30	0,001	1,275	89,315
3	0,003	5,421	25,759	17	0,001	2,069	68,145	31	0,001	1,218	90,533
4	0,003	4,498	30,257	18	0,001	2,032	70,176	32	0,001	1,179	91,712
5	0,003	4,235	34,492	19	0,001	1,919	72,096	33	0,001	1,122	92,834
6	0,002	4,020	38,512	20	0,001	1,855	73,951	34	0,001	1,119	93,953
7	0,002	3,732	42,243	21	0,001	1,806	75,757	35	0,001	1,008	94,962
8	0,002	3,347	45,590	22	0,001	1,729	77,486	36	0,001	0,949	95,910
9	0,002	3,138	48,728	23	0,001	1,687	79,173	37	0,001	0,870	96,781
10	0,002	2,899	51,627	24	0,001	1,601	80,774	38	0,001	0,855	97,636
11	0,002	2,708	54,335	25	0,001	1,564	82,339	39	0,000	0,754	98,390
12	0,002	2,483	56,818	26	0,001	1,489	83,828	40	0,000	0,650	99,040
13	0,001	2,421	59,240	27	0,001	1,475	85,303	41	0,000	0,549	99,589
14	0,001	2,367	61,606	28	0,001	1,390	86,693	42	0,000	0,411	100,000

Tabela 9: Autovalores em ordem decrescente para o segundo modelo KPCA, usado para fins de construção do sistema de detecção de falhas.

Também de modo similar ao passo 2 da Etapa 1 anterior de Construção e validação dos sistemas de detecção de falhas, a Tabela 10 mostra os limites superiores de controle para os modelos PCA, KPCA e DPCA, equivalentes ao percentil 99, ou seja, dada a taxa de alarmes falsos de 1% (conforme Russell et al., 2000).

Técnica	T^2	Q
PCA	54,386	24,749
KPCA	0,10896	5,8361e-05
DPCA	108,130	83,868

Tabela 10: Limites superiores de controle para T^2 e Q.

Na sequência, obtiveram-se os valores das métricas de monitoramento, T^2 e Q, a partir de cada modelo em separado, e calcularam-se então as métricas combinadas para T^2 e Q. Em seguida, calculou-se a métrica de monitoramento, denominada Taxa de detecções perdidas (*Missed Detection Rate*, MDR), relativas às métricas combinadas T^2 e Q. A Tabela 11 mostra os resultados relativos à T^2 , e a Tabela 12, à Q. Pode observar valores relativamente discrepantes de MDR, principalmente para a estatística Q, o que mostra uma capacidade relativamente diferente de detecção entre as técnicas, em que nenhuma delas apresenta o melhor desempenho ao se considerar todo o conjunto de falhas. Esse fato exemplifica a motivação para a adoção de um sistema combinado de detecção de falhas. Incluíram-se os resultados de Russell et al. (2000) obtidos com PCA apenas como referência, uma vez que o percentual de variância explicada acumulada é significativamente menor em relação aos modelos individuais.

Inicialmente, pode-se observar que o resultado de qualquer modelo em separado nunca é superior ao sistema combinado para além de 5%, sendo máximo para a técnica DPCA, falha 19 e métrica T^2 , com diferença em relação ao sistema combinado de -3,241% (em *itálico*). Esse resultado é importante para mostrar que não há perda de detecção com o sistema em relação àqueles individuais.

Por outro lado, pode-se observar um resultado superior do sistema combinado, com uma diferença acima de 5% em relação aos três modelos individuais, de modo simultâneo, para um conjunto de falhas. Em relação à métrica T^2 , observa-se para um conjunto significativo de sete falhas: 4, 5, 10, 11, 16, 17 e 20 (em **negrito**). Nesse caso, as menores diferenças percentuais (ainda sim, relativamente altas) são para as falhas 4 (14,410; 26,271; 14,125), 16 (21,148; 19,400; 21,125) e 17 (18,597; 16,725; 18,500), e as maiores, para as falhas 5 (75,531; 73,658; 75,500), 10 (29,670; 21,056; 29,625), 11 (51,809; 61,422; 51,750) e 20 (56,790; 48,800; 56,625). Em relação à métrica Q, as diferenças acima de 5% são para um conjunto também significativo de seis falhas: 5, 10, 11, 16, 19 e 20 (em **negrito**). Nesse caso, as diferenças percentuais são significativamente altas para todas as falhas, sendo, para a

falha 5 (76,279; 76,779; 78,500), falha 10 (39,674, 38,551; 61,875), falha 11 (48,938; 11,485; 49,500), falha 16 (48,042; 44,546; 71,125), falha 19 (38,638; 12,171; 44,000) e falha 20 (39,813; 32,322; 42,125).

A partir dos resultados obtidos por Russell et al. (2000) para a técnica PCA, pode-se dividir as falhas em quatro grupos, conforme o grau de dificuldade de detecção: Grupo 1 (detecção fácil; com MDR (*Missed Detection Rate*) < 0,1): Falhas 1, 2, 4, 6, 7, 8, 12, 13, 14; Grupo 2 (detecção intermediária; com $0,1 < \text{MDR} < 0,5$): Falhas 11, 17 e 18; Grupo 3 (detecção difícil; com $0,5 < \text{MDR} < 0,9$): Falhas 5, 10, 16, 19, 20 e 21; e Grupo 4 (detecção muito difícil; $\text{MDR} > 0,9$): Falhas 3, 9 e 15. Conforme a descrição anterior, o sistema combinado é capaz de melhorar, de modo significativo, a capacidade de detecção da falha 4 (grupo 1); das falhas 11 e 17 (grupo 2); e de cinco das seis falhas do grupo 3, a saber, falhas 5, 10, 16 e 20, tanto em relação a T^2 quanto a Q, e 19, em relação a Q. Pode-se verificar que o valor de MDR é significativamente próximo de zero para as falhas 5, 11, 17 e 20 (valores sublinhados), em relação à T^2 , e para as falhas 5, 11 e 20, em relação à Q.

A partir desse conjunto de resultados, é possível verificar a relevância de se empregar uma abordagem baseada em sistema combinado de detecção de falhas, em relação àqueles individuais.

Falha	Russell et al. (2000)	PCA k = 27	DPCA k = 67 l = 3	KPCA D = 42 k = 27	Ensemble métrica (EM)	Diferença (%) ($T_{Russell}^2 - T_{EM}^2$)	Diferença (%) ($T_{PCA}^2 - T_{EM}^2$)	Diferença (%) ($T_{DPCA}^2 - T_{EM}^2$)	Diferença (%) ($T_{KPCA}^2 - T_{EM}^2$)
1	0,008	0,007	0,007	0,006	0,006	0,175	0,124	0,124	0,000
2	0,020	0,019	0,016	0,018	0,016	0,375	0,248	-0,002	0,125
3	0,998	0,990	0,993	0,990	0,999	-0,075	-0,874	-0,624	-0,875
4	0,956	0,717	0,835	0,714	0,573	38,350	14,410	26,271	14,125
5	0,775	0,755	0,737	0,755	<u>0,000</u>	77,500	75,531	73,658	75,500
6	0,011	0,011	0,010	0,010	0,010	0,100	0,124	-0,001	0,000
7	0,085	0,001	0,001	0,000	0,000	8,500	0,125	0,125	0,000
8	0,034	0,029	0,029	0,028	0,028	0,650	0,121	0,121	0,000
9	0,994	0,989	0,994	0,989	0,995	-0,100	-0,624	-0,124	-0,625
10	0,666	0,638	0,552	0,638	0,341	32,475	29,670	21,056	29,625
11	0,794	0,527	0,623	0,526	<u>0,009</u>	78,525	51,809	61,422	51,750
12	0,029	0,016	0,010	0,015	0,003	2,650	1,373	0,749	1,250
13	0,060	0,051	0,050	0,050	0,048	1,250	0,369	0,244	0,250
14	0,158	0,002	0,002	0,001	0,000	15,800	0,250	0,250	0,125
15	0,988	0,978	0,956	0,978	0,979	0,925	-0,122	-2,245	-0,125
16	0,834	0,818	0,800	0,818	0,606	22,775	21,148	19,400	21,125
17	0,259	0,221	0,202	0,220	<u>0,035</u>	22,400	18,597	16,725	18,500
18	0,113	0,111	0,109	0,110	0,110	0,300	0,111	-0,139	0,000
19	0,996	0,956	0,925	0,956	0,958	3,850	-0,120	-3,241	-0,125
20	0,701	0,677	0,597	0,675	<u>0,109</u>	59,225	56,790	48,800	56,625
21	0,736	0,642	0,571	0,641	0,580	15,600	6,170	-0,946	6,125

Tabela 11: Taxa de detecções perdidas (MDR, *Missed Detection Rate*), relativas à métrica de monitoramento T^2 , para os modelos individuais (PCA, DPCA e KPCA) e aquele combinado (EL).

Falha	Russell et al. (2000)	PCA k = 27	DPCA k = 67 l = 3	KPCA D = 42 k = 27	Ensemble métrica (EM)	Diferença (%) ($Q_{Russell} - Q_{EM}$)	Diferença (%) ($Q_{PCA} - Q_{EM}$)	Diferença (%) ($Q_{DPCA} - Q_{EM}$)	Diferença (%) ($Q_{KPCA} - Q_{EM}$)
1	0,003	0,004	0,005	0,003	0,003	0,050	0,125	0,249	0,000
2	0,014	0,014	0,019	0,013	0,013	0,150	0,123	0,623	0,000
3	0,991	0,984	0,988	0,986	0,991	-0,025	-0,748	-0,373	-0,500
4	0,038	0,024	0,001	0,028	0,000	3,800	2,372	0,125	2,750
5	0,746	0,764	0,769	0,786	<u>0,001</u>	74,475	76,279	76,779	78,500
6	0	0,001	0,001	0,006	0,000	0,000	0,125	0,125	0,625
7	0	0,001	0,001	0,000	0,000	0,000	0,125	0,125	0,000
8	0,024	0,039	0,024	0,039	0,023	0,150	1,620	0,122	1,625
9	0,981	0,976	0,990	0,979	0,984	-0,275	-0,747	0,626	-0,500
10	0,659	0,604	0,593	0,826	0,208	45,150	39,674	38,551	61,875
11	0,356	0,497	0,122	0,503	<u>0,008</u>	34,850	48,938	11,485	49,500
12	0,025	0,041	0,015	0,056	0,003	2,250	3,870	1,248	5,375
13	0,045	0,052	0,054	0,051	0,049	-0,375	0,368	0,493	0,250
14	0	0,002	0,001	0,001	0,000	0,000	0,250	0,125	0,125
15	0,973	0,968	0,993	0,969	0,969	0,425	-0,121	2,376	0,000
16	0,755	0,667	0,632	0,898	0,186	56,875	48,042	44,546	71,125
17	0,108	0,070	0,030	0,074	0,026	8,175	4,366	0,371	4,750
18	0,101	0,094	0,099	0,093	0,093	0,850	0,113	0,613	0,000
19	0,873	0,896	0,632	0,950	0,510	36,300	38,638	12,171	44,000
20	0,550	0,498	0,423	0,521	<u>0,100</u>	45,000	39,813	32,322	42,125
21	0,570	0,553	0,578	0,558	0,553	1,750	0,056	2,553	0,500

Tabela 12: Taxa de detecções perdidas (MDR, *Missed Detection Rate*), relativas à métrica de monitoramento Q, para os modelos individuais (PCA, DPCA e KPCA) e aquele combinado (EL).

A Figura 8 mostra, de modo gráfico, os resultados apresentados nas Tabelas 11 e 12.

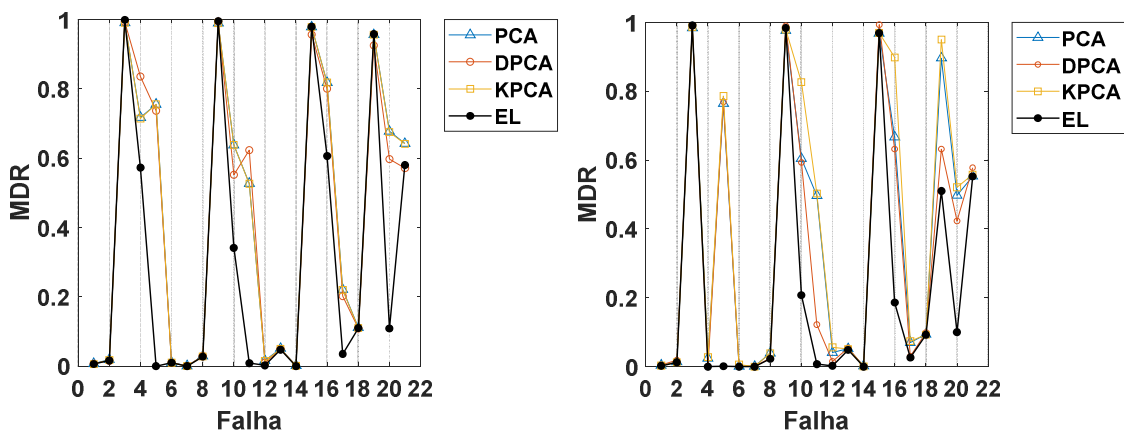
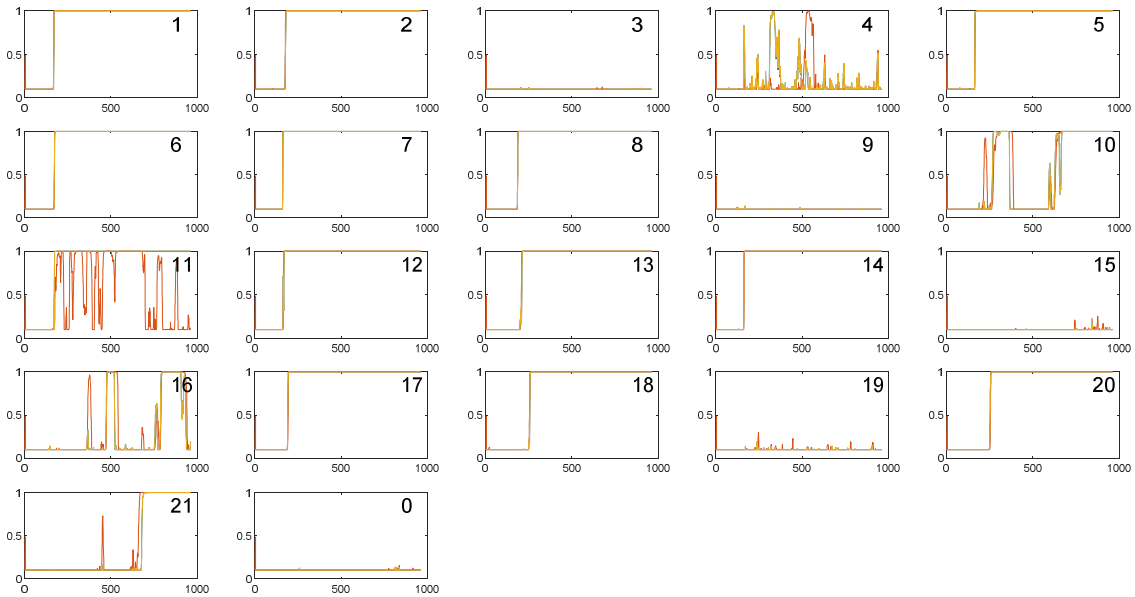
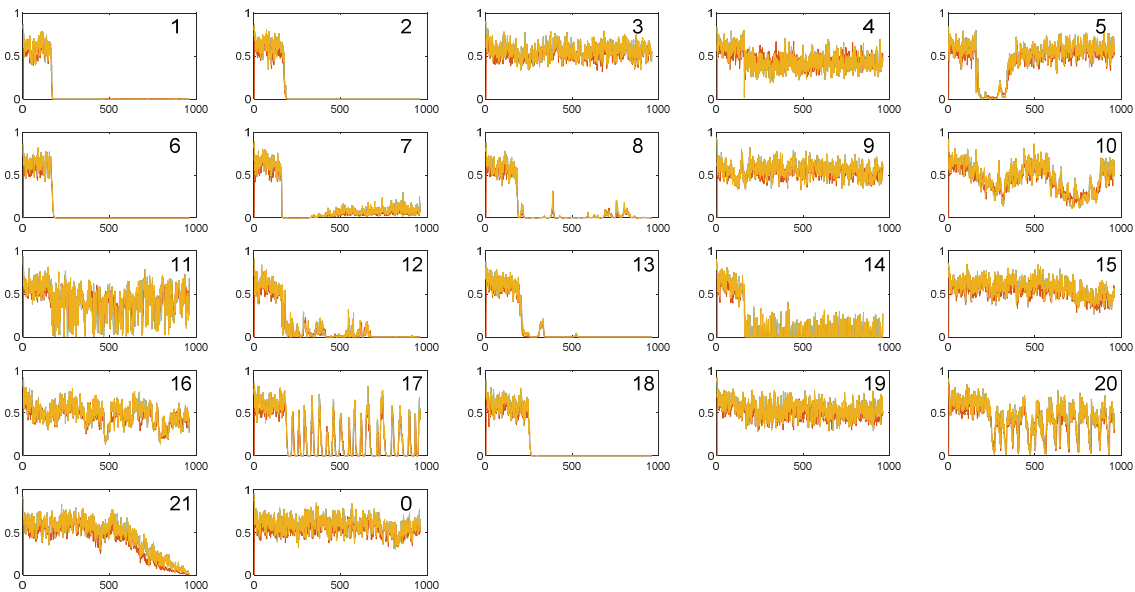


Figura 8: MDR (*Missed Detection Rate*), relativo às estatísticas (a) T^2 e (b) Q, para os sistemas individuais (PCA, DPCA e KPCA) e aquele combinado (EL).

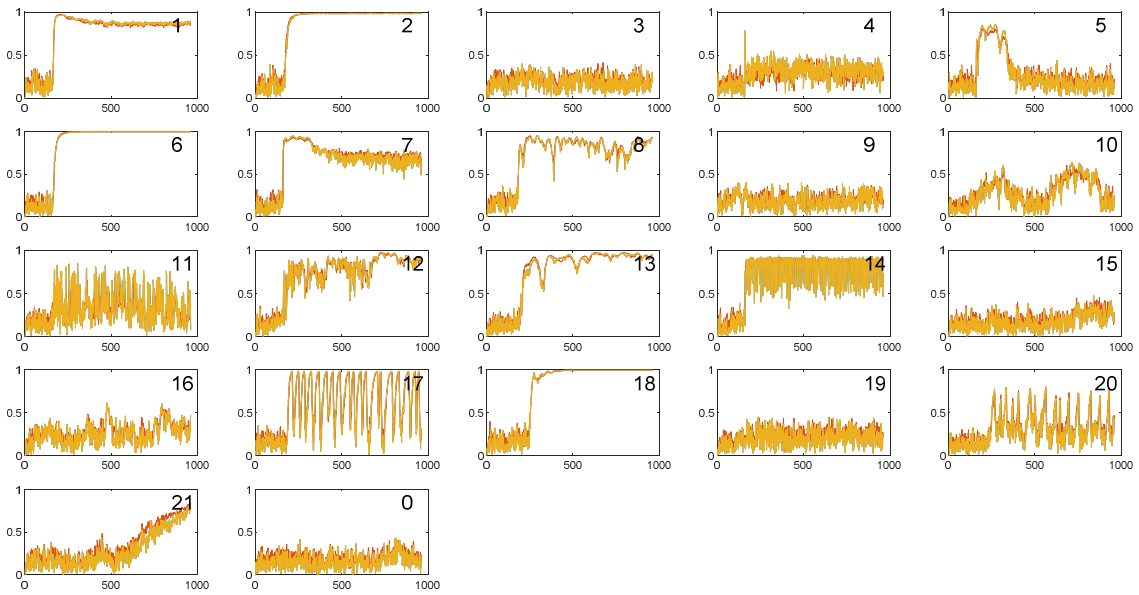
A Figura 9 apresenta os cálculos intermediários de probabilidade *a priori*, termos de verossimilhança e de probabilidade *a posteriori*, e o resultado final para a *ensemble* métrica, relativo à estatística T^2 . A Figura 10 apresenta os mesmos resultados; porém, relativos à estatística Q .



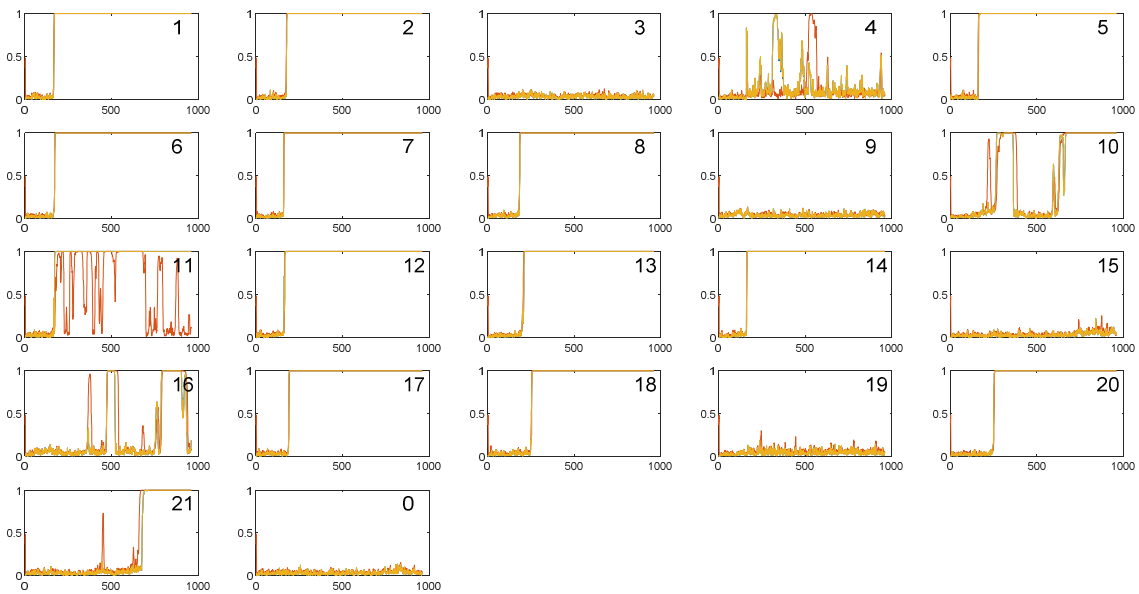
(a) Probabilidade *a priori* ($P(F)$) para os modelos PCA, KPCA e DPCA.



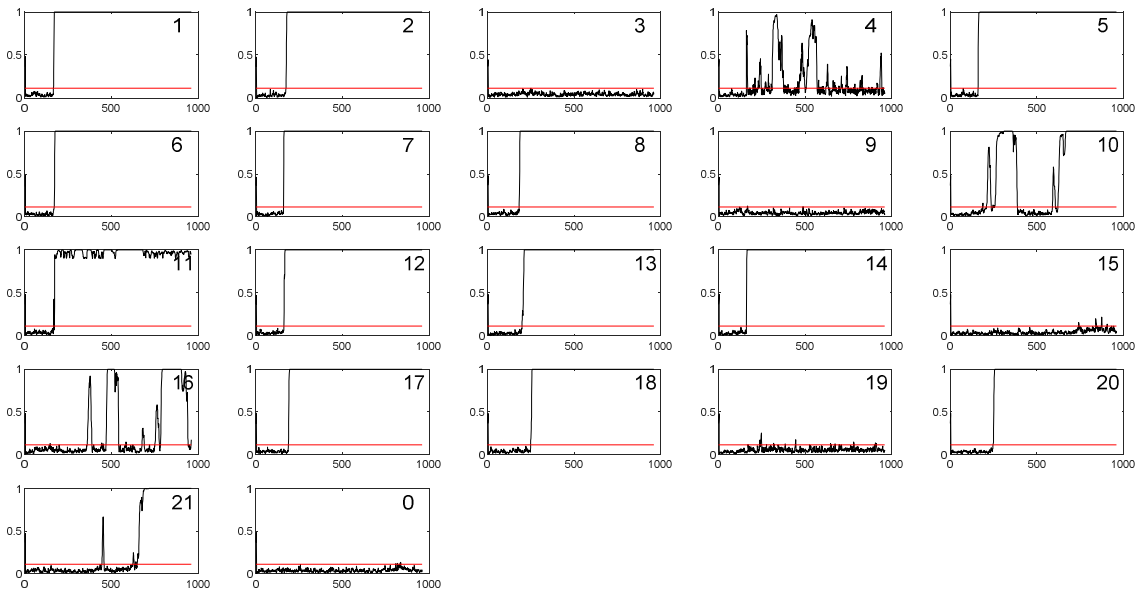
(b) Termo de verossimilhança ($P(x|N)$) para os modelos PCA, KPCA e DPCA.



(c) Termo de verossimilhança ($P(x|F)$) para os modelos PCA, KPCA e DPCA.

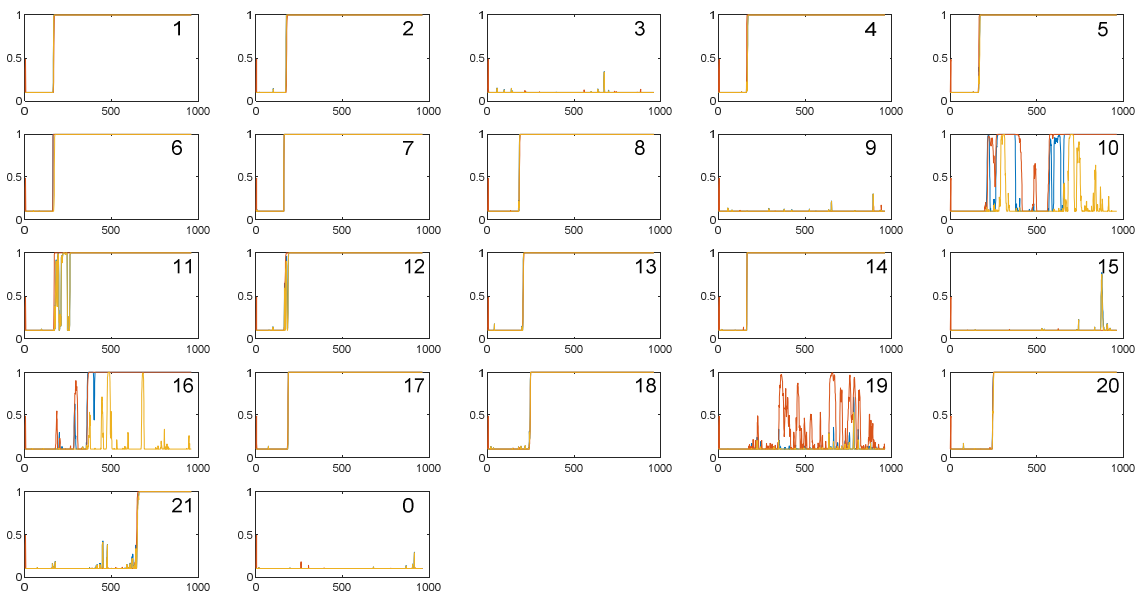


(d) Probabilidade *a posteriori* ($P(F|x)$) para os modelos PCA, KPCA e DPCA.

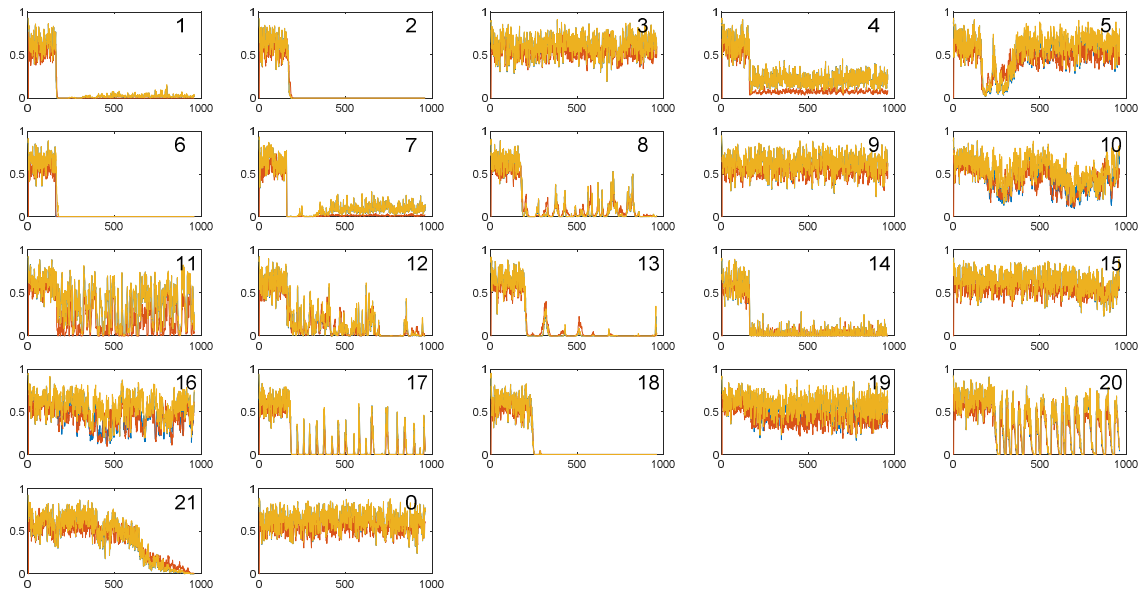


(e) *Ensemble métrica (ET^2).*

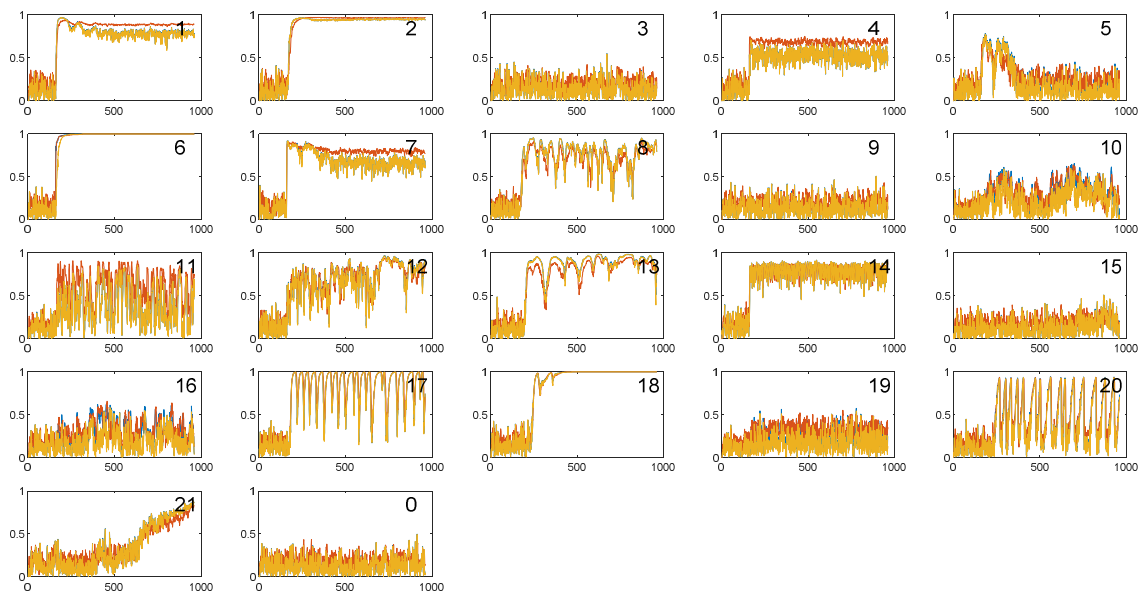
Figura 9: Resultados relativos à estatística T^2 , em que 1 refere-se à falha 1, 2 à falha 2, e assim por diante até a falha 21, e 0, à operação normal.



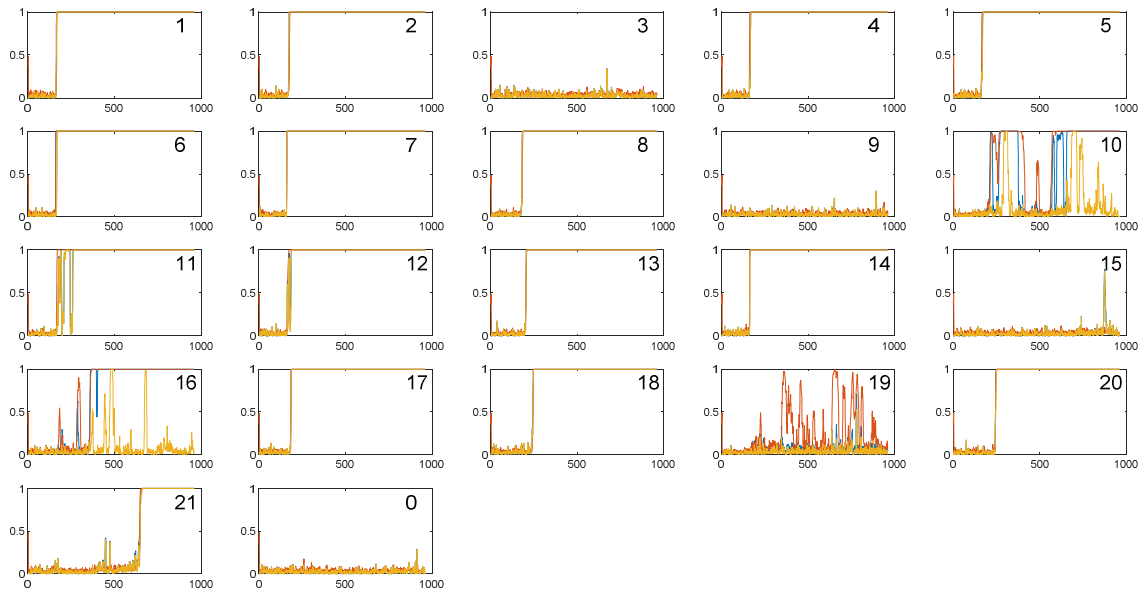
(a) *Probabilidade a priori ($P(F)$) para os modelos PCA, KPCA e DPCA.*



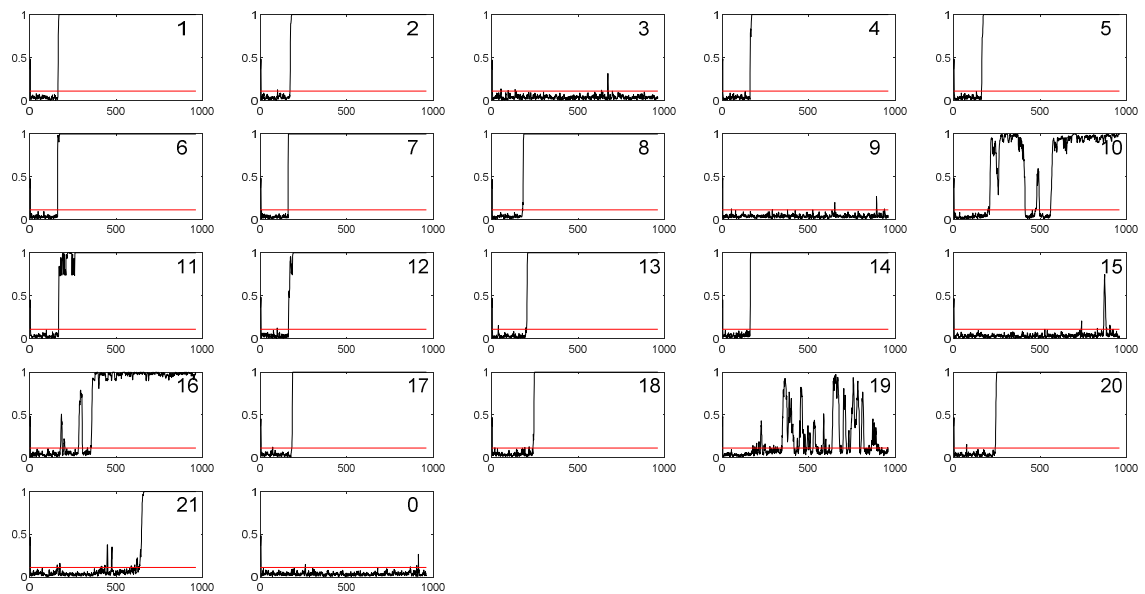
(b) Termo de verossimilhança ($P(x|N)$) para os modelos PCA, KPCA e DPCA.



(c) Termo de verossimilhança ($P(x|F)$) para os modelos PCA, KPCA e DPCA.



(d) Probabilidade *a posteriori* ($P(F|x)$) para os modelos PCA, KPCA e DPCA.



(e) *Ensemble métrica (EQ)*.

Figura 10: Resultados relativos à estatística Q , em que 1 refere-se à falha 1, 2 à falha 2, e assim por diante até a falha 21, e 0, à operação normal.

8. CONCLUSÕES

Inicialmente, destaca-se a capacidade dos modelos PCA, KPCA e DPCA, em explicar a maior parte da variância total dos dados originais, a partir de um número significativamente menor de componentes (variáveis latentes). Essa possibilidade é importante por facilitar a análise e interpretação de problemas multivariados.

Neste trabalho, combinaram-se os resultados de três técnicas estatísticas, com o objetivo de se alcançar um melhor desempenho em relação à detecção de falhas em processos químicos. A primeira é aquela denominada Análise por Componentes Principais (PCA), que é a técnica mais usual em aplicações de métodos estatísticos em monitoramento de processos químicos desde os anos 90. Dado o seu princípio de transformação linear, não é capaz de capturar a não linearidade, que é uma característica comum aos processos químicos. Esse fato é a motivação para a segunda técnica, denominada *kernel* PCA, ou KPCA. Outra característica em processos químicos é a auto-correlação em variáveis de processos. Assim, a terceira técnica é aquela denominada PCA dinâmico, ou DPCA.

Inicialmente, construiu-se e validou-se, em separado, um sistema de detecção baseado em cada uma dessas técnicas. Na sequência, construíram-se, ainda, sistemas individuais; porém, com uma maior capacidade de explicação da variância total das variáveis originais (85%). Nesse ponto, observaram-se capacidades de detecção dos modelos relativamente distintas, dado uma falha em particular. Esse fato evidencia que nenhuma técnica em particular é capaz de apresentar o melhor desempenho para todos os tipos de falhas, e que, portanto, a combinação desses resultados constituiu-se em uma solução alternativa em potencial para aumentar a eficiência de detecção de falhas. Nessa direção, combinaram-se os resultados para as estatísticas T^2 e Q dos três modelos a partir de uma abordagem de *ensemble learning* baseada em inferência Bayesiana.

A partir da análise dos resultados, observou-se que não há perda da capacidade de detecção do sistema combinado em relação àqueles individuais; a melhora significativa do desempenho, a partir da métrica taxa de detecções

perdas (*Missed Detection Rate*, MDR), principalmente para um grupo de falhas de difícil detecção, e valores de MDR consideravelmente próximos a zero para algumas dessas falhas. Em resumo, foi possível verificar o ganho significativo ao se adotar um sistema combinado de detecção de falhas em relação aos modelos individuais.

8.1. Sugestões de Trabalhos Futuros

Como sugestões de trabalho futuros, ressalta-se,

- a inclusão de outros métodos estatísticos multivariados, como por exemplo, a Análise de Componentes Independentes (*Independent Component Analysis*, ICA), útil para a descrição de processos não-gaussianos,
- a comparação da estratégia de *ensemble learning* baseada em inferência Bayesiana com outras abordagens,
- o desenvolvimento específico de estratégias de *ensemble learning* para as falhas classificadas com grau de detecção muito difícil,
- e o uso de outras métricas de monitoramento, além da Taxa de detecções perdidas, como por exemplo, o tempo até a detecção.

9. REFERÊNCIAS BIBLIOGRÁFICAS

- Amaral, L. M. O conceito de reindustrialização, indústria 4.0 e a política industrial para o século XXI – O caso português. Política industrial para o século XXI. Disponível em http://cip.org.pt/wp-content/uploads/2016/05/LUIS_MIRA_AMARAL.pdf, Acesso em 20 de Julho de 2017.
- ASM (Abnormal Situation Management) - A Joint Research and Development Consortium. Disponível em <http://www.asmconsortium.org/>, Acesso em 01 de Outubro de 2018.
- ASQ - American Society for Quality, The history of quality: Overview. Disponível em <http://asq.org/learn-about-quality/history-of-quality/overview/overview.html>, Acesso em 12 de Junho de 2017.
- Chiang, L. H., Russell, E. L., Braatz, R. D. Fault detection and diagnosis in industrial systems, Springer, London, 2001.
- Cho, J. H., Lee, J. M., Choi, S. W., Lee, D. W., Lee, I. B. Fault identification for process monitoring using kernel principal component analysis, Chemical Engineering Science, 60(1), 279-288, 2005.
- Control, A. S. Statistical process control manual, 1986.
- Cui, P., Li, J., Wang, G. Improved kernel principal component analysis for fault detection, Expert Systems with Applications, 34(2), 1210-1219, 2008.
- Das, A., Maiti, J., Banerjee, R. N. Process monitoring and fault detection strategies: A review. International Journal of Quality & Reliability Management, 29(7), 720-752, 2012.
- Dong, D., McAvoy, T. J. Nonlinear principal component analysis based on principal curves and neural networks. Computers and Chemical Engineering, 20, 65-78, 1996.
- Downs, J. J., Vogel, E. F. A plant-wide industrial process control problem, Computers and Chemical Engineering, 17(3), 245-255, 1993.
- Du, Y., Du, D. Fault detection and diagnosis using empirical mode decomposition based principal component analysis, Computers & Chemical Engineering, 115(12), 1-21, 2018.

- Ge, Z., Song, Z. Multivariate statistical process control, process monitoring methods and applications, 45-80, Springer, 2013.
- Ge, Z., Song, Z., Gao, F. Review of recent research on data-based process monitoring. *Industrial and Engineering Chemical Research*, 52, 3543–3562, 2013.
- Huang, J., Yan, X. Dynamic process fault detection and diagnosis based on dynamic principal component analysis, dynamic independent component analysis and Bayesian inference, *Chemometrics and Intelligent Laboratory Systems*, 148, 115-127, 2015.
- INMETRO, Livro de Qualidade. Disponível em http://www.inmetro.gov.br/barreirastecnicas/pdf/Livro_Qualidade.pdf, Acesso em 12 de Junho de 2017.
- Jolliffe, I. T. *Principal component analysis*. Springer, 1986.
- Kahneman, D., Lovallo, D., Sibony, O. Before you make that big decision. *Harvard Business Review*, 89(6), 50-60, 2011.
- Kallas, M., Mouro, G., Anani, K., Ragot, J., Maquin, D. Fault detection and estimation using kernel principal component analysis, *IFAC-PapersOnLine*, 50(1), 1025-1030, 2017.
- Ku, W., Storer, R. H., Georgakis C. Disturbance detection and isolation by dynamic principal components analysis. *Chemometrics and Intelligent Laboratory Systems*, 30, 179-196, 1995.
- Kuncheva, L. I.; Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 2003, 51 (2), 181–207.
- Li N, Yang Y. Ensemble kernel principal component analysis for improved nonlinear process monitoring. *Industrial & Engineering Chemistry Research*, 54(1), 318-329, 2015.
- Mingoti, S. A., *Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada*, Ed. UFMG, 2007.
- Montgomery D. C., *Introduction to Statistical Quality Control*, 6thed., Wiley, 2008.
- Montgomery, D. C., Runger, G. C. *Estatística Aplicada e Probabilidade para Engenheiros*. ed. 4, LTC, 2009.
- Nguyen, V. H., Golival, J.-C. Fault detection based on kernel principal

- component analysis, *Engineering Structures*, 32(11), 3683-3691, 2010.
- Qin, S. J. Statistical process monitoring: Basics and beyond, *Journal of Chemometrics*, 17(8), 480-502, 2003.
 - Russell E. L., Chiang, L. H., Braatz, R. D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 51, 81-93, 2000.
 - Scholkopf, B., Smola, A., Muller, K. R. Kernel principal component analysis. In: Schölkopf, B., Burges, C. J. C., Smola, A. J. (eds), *Advances in kernel methods-Support vector learning*, MIT Press, 327-352, 1999.
 - Scholkopf, B., Smola, A., Muller, K. R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computing*, 10(5), 1299-1319, 1998.
 - Scholz, M., Kaplan F., Guy, C. L., Kopka, J., Selbig, J., Non-linear PCA: A missing data approach. *Bioinformatics*, 21(15), 3887-3895, 2005.
 - Severson, P., Chaiwatanodom, P., Braatz, R. D. Perspectives on process monitoring of industrial systems. *Annual Reviews in Control*, 42, 190–200, 2016.
 - Van der Maaten, L. *Matlab Toolbox for DimensionalityReduction*, Disponível em <https://lvdmaaten.github.io/drtoolbox/>, Acesso em 01 de Agosto de 2018.
 - Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N. A review of process fault diagnosis - Part II: Qualitative models and search strategies. *Computers and Chemical Engineering*, 27(3), 313-326, 2003b.
 - Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K. A review of process fault diagnosis - Part III: Process history based methods. *Computers and Chemical Engineering*, 27(3), 327-346, 2003c.
 - Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N. A review of process fault diagnosis - Part I: Quantitative model-based methods. *Computers and Chemical Engineering*, 27(3), 293-311, 2003a.
 - Wang, Z., Wang, L., Liang, K., Tan, Y. Enhanced chiller fault detection using Bayesian network and principal component analysis, *Applied Thermal Engineering*, 141, 898-905, 2018.
 - Wise, M. Adapting multivariate analysis for monitoring and modeling of dynamic systems. Doctoral dissertation, University of Washington, 1991.
 - Yin, S. Ding, S. X., Xie, X. Luo, H. A Review on basic data-driven

approaches for industrial process monitoring. IEEE Transactions on Industrial Electronics, 61(11), 6418-6428, 2014.

- Yu, Z, Wang D, You J., et al. Progressive subspace ensemble learning. Pattern Recognition, 60, 692-705, 2016.