UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Fernanda Araujo Pimentel Peres

SELEÇÃO DE VARIÁVEIS APLICADA AO CONTROLE ESTATÍSTICO MULTIVARIADO DE PROCESSOS EM BATELADAS

Porto Alegre

2018

Fernanda Araujo Pimentel Peres

Seleção de Variáveis Aplicada ao Controle Estatístico Multivariado de Processos em Bateladas

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutor em Engenharia, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientador: Flávio Sanson Fogliatto, *Ph.*D.

Porto Alegre

2018

Fernanda Araujo Pimentel Peres

Seleção de Variáveis Aplicada ao Controle Estatístico Multivariado de Processos em Bateladas

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

Prof. Flávio Sanson Fogliatto, *Ph.*D.

Orientador, PPGEP/UFRGS

Prof. Michel José Anzanello, *Ph.*D.

Vice-Coordenador, PPGEP/UFRGS

Banca Examinadora:

Professor André Luis Korzenowski, Dr. (PPGEPS / UNISINOS)

Professor Marcelo Farenzena, Dr. (PPGEQ / UFRGS)

Professor Michel José Anzanello, *Ph.D.* (PPGEP / UFRGS)

Ao meu marido Thiago

e aos meus pais Denise e Fernando

AGRADECIMENTOS

Ao meu marido Thiago pelo suporte emocional e técnico, pelas infinitas horas de programações, por ter acreditado em mim em todos os momentos e ser meu maior incentivador. Nada disso seria possível sem ti. Te amo!

Aos meus amados pais Denise e Fernando por me estimularem, serem meu porto seguro, exemplo de vida e respeito, estarem sempre presentes e por todo amor dedicado.

Ao meu orientador Flávio pelo exemplo de profissional, pelo apoio e dedicação dispensados, pela disponibilidade constante para sanar minhas dúvidas e por todo o conhecimento transmitido. Muito obrigada!

Ao professor Michel, por todo o conhecimento partilhado e disponibilidade em me auxiliar ao longo do andamento deste trabalho.

Ao meu irmão Rafael pelo apoio nos momentos difíceis e ao meu irmão Guilherme por todos os almoços, dicas e orientações e por sempre me estimular a alçar voos maiores. À eles e minhas cunhadas, Fabrícia e Bruna, que me deram as mais lindas sobrinhas Laura e Isabela e a honra de ser a tia-dinda da amada Luíza.

À minha amiga-irmã Fabiana por saber tudo de mim, sempre me acolher, crescer comigo e estar junto em todos os momentos.

Aos meus amigos do mundo dos chocolates, Elisângela, Mallmann, Flávio, Fernanda e Bruno, que me abriram as portas, compartilharam conhecimentos e sempre me receberam com cordialidade e amizade, contribuindo muito para este trabalho.

Aos amigos que tenho a honra de conviver (Choco Amigas, Sempre Unidos e Sulas) que foram fundamentais neste período de muito trabalho, sempre trazendo alegria e leveza para minha vida. Ao pessoal do Creties pela convivência e aos ‘Amigos Sem Fronteiras’, Caroline, Érica, Giuliana, Letícia, Marlon, Ricardo e Vítor, que eu tive o prazer de conhecer no PPGEP, foram fundamentais para o “suporte psicológico” ao longo desta jornada e os quais eu espero levar para minha vida.

À UFRGS, Universidade pública gratuita e de qualidade que me acolheu nos últimos 20 anos fazendo de mim a profissional que eu sou hoje. E ao CNPq que me garantiu o suporte financeiro para me dedicar à execução deste trabalho.

À minha terapeuta Elenara por me auxiliar a me conhecer, crescer e manter meu equilíbrio emocional mesmo nos momentos mais difíceis.

E a Santo Antônio, por atender às minhas preces e me dar suporte espiritual para seguir em frente.

‘‘Feliz aquele que transfere o que sabe e aprende o que ensina.’’

Cora Coralina

PERES, Fernanda Araujo Pimentel. ***Seleção de Variáveis Aplicada ao Controle Estatístico Multivariado de Processos em Bateladas***, 2018. Tese (Doutorado em Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

A presente tese apresenta proposições para o uso da seleção de variáveis no aprimoramento do controle estatístico de processos multivariados (MSPC) em bateladas, a fim de contribuir com a melhoria da qualidade de processos industriais. Dessa forma, os objetivos desta tese são: (*i*) identificar as limitações encontradas pelos métodos MSPC no monitoramento de processos industriais; (*ii*) entender como métodos de seleção de variáveis são integrados para promover a melhoria do monitoramento de processos de elevada dimensionalidade; (*iii*) discutir sobre métodos para alinhamento e sincronização de bateladas aplicados a processos com diferentes durações; (*iv*) definir o método de alinhamento e sincronização mais adequado para o tratamento de dados de bateladas, visando aprimorar a construção do modelo de monitoramento na Fase I do controle estatístico de processo; (*v*) propor a seleção de variáveis, com propósito de classificação, prévia à construção das cartas de controle multivariadas (CCM) baseadas na análise de componentes principais (PCA) para monitorar um processo em bateladas; e (*vi*) validar o desempenho de detecção de falhas da carta de controle multivariada proposta em comparação às cartas tradicionais T^2 e Q baseadas em PCA. O desempenho do método proposto foi avaliado mediante aplicação em um estudo de caso com dados reais de um processo industrial alimentício. Os resultados obtidos demonstraram que a realização de uma seleção de variáveis prévia à construção das CCM contribuiu para reduzir eficientemente o número de variáveis a serem analisadas e superar as limitações encontradas na detecção de falhas quando bancos de elevada dimensionalidade são monitorados. Conclui-se que, ao possibilitar que CCM, amplamente utilizadas no meio industrial, sejam adequadas para banco de dados reais de elevada dimensionalidade, o método proposto agrega inovação à área de monitoramento de processos em bateladas e contribui para a geração de produtos de elevado padrão de qualidade.

Palavras-chave: Seleção de variáveis. Controle estatístico de processos multivariados.
Processo em bateladas. Detecção de falhas.

PERES, Fernanda Araujo Pimentel. ***Variable Selection Applied to Multivariate Statistical Control of Batch Processes***, 2018. Dissertation (Doctorate in Industrial Engineering) – Universidade Federal do Rio Grande do Sul, Brazil.

ABSTRACT

This dissertation presents propositions for the use of variable selection in the improvement of multivariate statistical process control (MSPC) of batch processes, in order to contribute to the enhacement of industrial processes' quality. There are six objectives: (*i*) identify MSPC limitations in industrial processes monitoring; (*ii*) understand how methods of variable selection are used to improve high dimensional processes monitoring; (*iii*) discuss about methods for alignment and synchronization of batches with different durations; (*iv*) define the most adequate alignment and synchronization method for batch data treatment, aiming to improve Phase I of process monitoring; (*v*) propose variable selection for classification prior to establishing multivariate control charts (MCC) based on principal component analysis (PCA) to monitor a batch process; and (*vi*) validate fault detection performance of the proposed MCC in comparison with traditional PCA-based T^2 and Q charts. The performance of the proposed method was evaluated in a case study using real data from an industrial food process. Results showed that performing variable selection prior to establishing MCC contributed to efficiently reduce the number of variables and overcome limitations found in fault detection when high dimensional datasets are monitored. We conclude that by improving control charts widely used in industry to accomodate high dimensional datasets the proposed method adds innovation to the area of batch process monitoring and contributes to the generation of high quality standard products.

Keywords: Variable selection. Multivariate statistical process control. Batch processes. Fault detection.

LISTA DE FIGURAS

Figura 2. 1 Search steps and results 31

Figura 2. 2 Interaction between VS and MSPC steps in class (a.1) Preprocessing Filter Approach, (a.2) Postprocessing Filter Approach, and (b) Wrapper Approach. 39

Figura 2. 3 VS-MSPC integration approaches clustered according to objectives and step of process monitoring they address 58

Figura 3. 1 Scaled pre-treated variables before synchronization. 81

Figura 3. 2 Variables' profiles after synchronization using (a) KMT (top), (b) RSWS (middle), and (c) GFW (bottom) methods 83

Figura 3. 3 Percentage of total weight per iteration number by (a) KMT (top), (b) RSWS (middle), and (c) GFW methods (bottom) 85

Figura 4. 1 Unfolding of the three-dimensional process data matrix 95

Figura 4. 2 PVS-MPCA method – implementation steps 103

Figura 4. 3 Average accuracy classification for different candidate sets of variables . 106

Figura 4. 4 Final iteration of Phase I of the MPCA method 108

Figura 4. 5 Monitoring of new batches in Phase II using the MPCA method 108

Figura 4. 6 Final iteration of Phase I using the PVS-MPCA method 109

Figura 4. 7 Monitoring of new batches in Phase II using the PVS-MPCA method. Green circle indicates errors according to industry classification 109

Figura 4. 8 Aligned and synchronized trajectories of original (unfolded) variables 113

LISTA DE TABELAS

Tabela 1. 1 Estrutura das etapas da pesquisa desenvolvida... 20

Tabela 2. 1 Acronyms used in the article 29

Tabela 2. 2 Evolution by year, journal title and country of origin of selected articles .. 32

Tabela 2. 3 Limitations of MSPC methods and specific objectives of articles covered in this review 34

Tabela 2. 4 Main characteristics of methods reviewed. 40

Tabela 3. 1 Classification accuracy, sensitivity, and specificity for synchronized batches; all values given in percentage... 87

Tabela 4. 1 Process variables measured during batch progression............................... 105

Tabela 4. 2 List of retained process variables and their time of measurement during batch progression.. 107

SUMÁRIO

1 INTRODUÇÃO ... 13

1.1 TEMA DA TESE ... 16

1.2 OBJETIVO DA TESE ... 16

1.3 JUSTIFICATIVA DO TEMA E OBJETIVOS ... 17

1.4 DELINEAMENTO DO ESTUDO ... 18

1.4.1 Método de Pesquisa ... 19

1.4.2 Método de Trabalho ... 19

1.5 DELIMITAÇÕES DO ESTUDO ... 22

1.6 ESTRUTURA DA TESE ... 23

1.7 REFERÊNCIAS ... 23

2 ARTIGO 1 – VARIABLE SELECTION METHODS IN MULTIVARIATE STATISTICAL PROCESS CONTROL: A SYSTEMATIC LITERATURE REVIEW ... 27

2.1 INTRODUCTION ... 27

2.2 METHOD ... 30

2.3 STUDY CHARACTERIZATION ... 31

2.4 PROPOSED VS-MSPC INTEGRATION METHODS ... 38

2.4.1 Preprocessing filter approach ... 43

2.4.2 Postprocessing filter approach ... 49

2.4.3 Wrapper approach ... 50

2.5 PROCESS MONITORING IN SPC AND PERFORMANCE OF THE DEVELOPED METHODS ... 56

2.5.1 Exclusive monitoring of potential OOC variables ... 59

2.5.2 Better modeling and prediction of response variables ... 62

2.5.3 Improvement in the monitoring of IC variables ... 62

2.6 CONCLUSION AND OPEN ISSUES ... 67

2.7 REFERENCES ... 69

3 ARTIGO 2 - STRATEGIES FOR SYNCHRONIZING CHOCOLATE CONCHING BATCH PROCESS DATA USING DYNAMIC TIME WARPING 73

3.1 INTRODUCTION ... 73

3.2 MATERIALS AND METHODS 76

3.2.1 Chocolate conching dataset 76

3.2.2 Pre-treatment of data 77

3.2.3 Alignment and synchronization of multivariate batch datasets 78

3.2.4 *k*-Nearest Neighbor classification method 79

3.3 RESULTS AND DISCUSSION.. 80

3.4 CONCLUSION 88

3.5 REFERENCES 88

4 ARTIGO 3 – FAULT DETECTION IN BATCH PROCESSES THROUGH VARIABLE SELECTION INTEGRATED TO MULTIWAY PRINCIPAL COMPONENT ANALYSIS 91

4.1 INTRODUCTION 91

4.2 BACKGROUND 94

4.2.1 Fault Detection using Batch-wise Multiway Principal Component Analysis (MPCA).. 94

4.2.2 Variable selection based on accuracy classification of batches 99

4.3 PVS-MPCA FAULT DETECTION METHOD.. 101

4.4 CASE STUDY AND DISCUSSION 104

4.4.1 Process description 104

4.4.2 Variable selection.. 105

4.4.3 Offline fault detection... 107

4.4.4 Discussion 110

4.5 CONCLUSION 114

4.6 REFERENCES 115

5 CONSIDERAÇÕES FINAIS.. 119

5.1 CONCLUSÕES 119

5.2 SUGESTÕES PARA TRABALHOS FUTUROS 122

5.3 REFERÊNCIAS 122

1 INTRODUÇÃO

O conceito principal do monitoramento de processos em bateladas é modelar as causas mais importantes de variação presentes sob condições normais de operação (VAN SPRANG et al., 2002). Caso variações entre bateladas (devidas a desvios das variáveis de processo das suas trajetórias específicas, erros no carregamento da receita e falhas potenciais da planta) não sejam detectadas ou corrigidas, pode ocorrer a produção de uma, ou de uma sequência de bateladas, de qualidade inconsistente. Tendo em vista a elevada competitividade dos mercados, a minimização dos custos decorrentes da má qualidade se torna mandatória (MARTIN; MORRIS; KIPARISSIDES, 1999; NOMIKOS; MACGREGOR, 1994, 1995a).

O monitoramento de um processo envolve duas fases. Na Fase I os dados são coletados com o objetivo de se adquirir conhecimento sobre o processo. Devem ser verificados dados não usuais, bem como a estabilidade do processo, de forma a desenvolver um modelo de monitoramento sob controle apropriado para ser utilizado na Fase II (WOODALL; MONTGOMERY, 2014). A Fase II, por sua vez, constitui-se de quatro etapas: detecção, isolamento e diagnóstico de falhas, e intervenção no processo. Na detecção de falhas, os comportamentos anormais do processo são reconhecidos; variáveis que mais contribuem para a falha detectada são isoladas e o diagnóstico de falhas determina as causas-raiz para ocorrência do sinal fora do controle. Por fim, a intervenção é conduzida para que os efeitos das falhas sejam removidos do processo e não mais gerem produtos em desacordo com a especificação (CHIANG; KOTANCHEK; KORDON, 2004; YAN; YAO, 2015).

Para o desempenho efetivo de um monitoramento, medidas das variáveis de processo (**X**) e das variáveis de qualidade final (**y**) devem ser obtidas. Uma dificuldade encontrada para processos em bateladas reside no fato desses dados serem altamente colineares e auto-correlacionados, podendo também existir dados faltantes (MACGREGOR et al., 1994). Tal complexidade na estrutura de correlação pode prejudicar a correta classificação das bateladas em conformes (de acordo com a especificação) ou não conformes (em desacordo com a especificação) (YAN; KUANG; YAO, 2017).

Para superar esses problemas, abordagens de controle estatístico de processo multivariado (MSPC ou *multivariate statistical process control)* baseadas em métodos de projeção como a análise de componentes principais (PCA ou *principal component* *analysis*) e a regressão por mínimos quadrados parciais (PLS ou *partial least squares regression*) foram desenvolvidas, para promover a redução da dimensionalidade do espaço de monitoramento a poucas variáveis latentes (KOURTI; MACGREGOR, 1995; MACGREGOR et al., 1994). Para análise de dados multivariados de processos em bateladas, variantes desses métodos devem ser utilizadas, como o PCA multidirecional (MPCA ou *multiway PCA*) (NOMIKOS; MACGREGOR, 1994) e o PLS multidirecional (MPLS ou *multiway PLS*) (NOMIKOS; MACGREGOR, 1995b). Em tais métodos a matriz tridimensional de dados **X**, de dimensão $(I \times J \times K)$, na qual I bateladas têm as trajetórias de suas J variáveis medidas em K intervalos de tempo, é desdobrada em uma matriz bidimensional de dimensão $(I \times JK)$. Assim se torna possível a análise da variabilidade existente entre bateladas em **X**, ao resumir a informação contida nos dados com relação a variáveis e a sua evolução no tempo.

Nas últimas décadas, a ampla disseminação das redes de sensores e dos sistemas de controle distribuídos contribuiu para a redução significativa dos custos e dificuldades relacionadas à coleta e armazenamento de informações. Dessa forma, bancos de dados de elevada dimensionalidade passaram a ser disponibilizados, compostos por centenas de medições de variáveis de processo (JIANG; YAN; HUANG, 2016; MEGAHED; JONES-FARMER, 2013; WOODALL; MONTGOMERY, 2014). Esse aumento na dimensionalidade das bases de dados fez com que a capacidade de detectar uma falha rapidamente e a habilidade de localizar variáveis que se deslocam se tornassem os grandes desafios do MSPC (JIANG; WANG; TSUNG, 2012; KUANG; YAN; YAO, 2015; WOODALL; MONTGOMERY, 2014). Sendo assim, o desenvolvimento de novos modelos estatísticos de controle de processo, como a integração da seleção de variáveis a métodos de MSPC para lidar com bancos de dados de elevada dimensionalidade de processos em bateladas, surge como um tópico promissor (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012; BISGAARD, 2012; JIANG; YAN; HUANG, 2016).

O principal objetivo dos métodos de seleção de variáveis é identificar o subconjunto de variáveis que carrega a informação mais relevante contida no conjunto completo de dados (ANZANELLO; FOGLIATTO, 2014). A melhoria do monitoramento através de cartas de controle multivariadas (CCM) pela integração com seleção de variáveis foi discutida por Capizzi (2015), que revisou a eficiência e as vantagens da abordagem combinada no monitoramento de somente um subconjunto de variáveis potencialmente responsáveis pelo alarme fora de controle. Recentemente, Peres e Fogliatto (2018) atualizaram e ampliaram o escopo daquele estudo ao apresentar uma revisão sistemática sobre a integração de métodos de seleção de variáveis com métodos de MSPC, abordando não somente o uso de cartas de controle para monitorar variáveis fora do controle, mas também os diversos *frameworks* utilizados para a seleção de variáveis de processo (sob controle ou em falha) com objetivo de aprimorar o monitoramento de processos multivariados. O uso de conhecimento de especialistas (ZARZO; FERRER, 2004), máquina de vetores de suporte (SVM ou *support vector machine)* (CHU; QIN; HAN, 2004) e algoritmos genéticos (GHOSH; RAMTEKE; SRINIVASAN, 2014) para selecionar variáveis com o objetivo de melhorar a detecção de falhas durante o monitoramento de processos multivariados são outros exemplos dessa aplicação. Peres e Fogliatto (2018) também destacaram a importância do desenvolvimento de pesquisas que combinem métodos de seleção de variáveis com matrizes de dados tridimensionais para eficientemente promover o diagnóstico em processos em bateladas. A identificação do grupo de variáveis que fornecem uma melhor classificação de bateladas leva ao aprimoramento do monitoramento de processo.

Dado esse contexto, surgem as questões de pesquisa que norteiam a presente tese. Em primeiro lugar: (*i*) qual o *status* atual dos métodos que integram a seleção de variáveis com métodos de MSPC, e quais limitações os mesmos se propõem a superar? Em segundo lugar, dado que a fabricação em bateladas apresenta variações no tempo de duração do processo, surge a questão: (*ii*) qual o método de sincronização e alinhamento de dados de bateladas é mais adequado, visando ao aprimoramento da construção do modelo de monitoramento na Fase I? Finalmente, (*iii*) é possível melhorar o desempenho na detecção de falhas quando bancos de dados de processos em bateladas de elevada dimensionalidade são monitorados por um método que integra o MSPC com a seleção de variáveis? Somando-se a isto, verifica-se que a melhoria de métodos de MSPC para processos em bateladas é bastante restrita e pouco explorada na literatura recente. A partir dessas observações, a presente tese visa a aprofundar o estudo dessas questões e propõe o desenvolvimento de um método a ser aplicado em um estudo de caso com dados industriais reais.

1.1 TEMA DA TESE

De acordo com a contextualização apresentada previamente, esta proposta de tese tem seu foco em mitigar as dificuldades de detecção de falhas quando um processo multivariado em bateladas de elevada dimensionalidade é monitorado. Assim, a inserção de uma etapa de seleção de variáveis prévia a elaboração das CCM através do procedimento de monitoramento multivariado desenvolvido por Nomikos e MacGregor (1994) é proposta. Almeja-se, desta forma, reduzir a probabilidade de alarmes falsos na detecção de falhas e, consequentemente, minimizar o número de variáveis a serem isoladas após um evento especial ser detectado.

Nesta tese, entende-se por detecção de falhas o momento no qual um sinal fora do controle é emitido pelas CCM T^2 ou Q baseadas em métodos de projeção (KOURTI; NOMIKOS; MACGREGOR, 1995; NOMIKOS; MACGREGOR, 1994; WOODALL; MONTGOMERY, 2014).

Os processos em bateladas são caracterizados pela sua flexibilidade, duração finita, comportamento não-linear, estado não estacionário, duração de processo variável e tempos diferentes de duração de eventos-chave entre bateladas (GARCÍA-MUÑOZ et al., 2003; MARTIN; MORRIS; KIPARISSIDES, 1999; NOMIKOS; MACGREGOR, 1995a). Esses são amplamente utilizados por indústrias químicas, farmacêuticas e de alimentos (GONZÁLEZ-MARTÍNEZ; FERRER; WESTERHUIS, 2011; KOURTI; NOMIKOS; MACGREGOR, 1995; RAMAKER et al., 2003) e resultam em bancos de dados de processo compostos por dezenas ou centenas de variáveis, tais como temperatura, pressão e concentração (MEGAHED; JONES-FARMER, 2013; NOMIKOS; MACGREGOR, 1995b; WANG; JIANG, 2009).

Finalmente, métodos de seleção de variáveis são considerados aqueles que identificam o subconjunto de variáveis que carrega a informação mais relevante contida no conjunto completo de dados. Exemplos incluem técnicas de seleção *forward* e *backward*, ferramentas de mineração de dados e ferramentas de otimização, como algoritmos genéticos (ANZANELLO; FOGLIATTO, 2014).

1.2 OBJETIVO DA TESE

O objetivo geral desta tese é propor um método que integre a seleção de variáveis ao controle estatístico de processo multivariado para aprimorar a detecção de falhas em bancos de dados de elevada dimensionalidade, oriundos de processos industriais em bateladas de duração variável.

Para que seja possível alcançar o objetivo geral deste trabalho, é necessário atingir os seguintes objetivos específicos:

a) Identificar as limitações encontradas pelos métodos MSPC no monitoramento de processos industriais.
b) Entender como métodos de seleção de variáveis são integrados para promover a melhoria do monitoramento de processos de elevada dimensionalidade.
c) Discutir sobre métodos para alinhamento e sincronização de bateladas aplicados a processos com diferentes durações.
d) Definir o método de alinhamento e sincronização mais adequado para o tratamento de dados de bateladas, visando a aprimorar a construção do modelo de monitoramento na Fase I do SPC.
e) Propor a seleção de variáveis, com propósito de classificação, prévia à construção das CCM baseadas em PCA para monitorar um processo em bateladas.
f) Validar o desempenho de detecção de falhas da carta de controle multivariada proposta em comparação às cartas tradicionais T^2 e Q baseadas em PCA.

1.3 JUSTIFICATIVA DO TEMA E OBJETIVOS

O tema desta tese envolve 3 áreas principais: (*i*) controle de processos multivariados de elevada dimensionalidade, (*ii*) seleção de variáveis integrada ao MSPC e (*iii*) processos industriais em bateladas. O aprimoramento do MSPC tem recebido destaque nos últimos anos. Tradicionalmente, a detecção de falhas baseada em métodos de projeção ocorre através de um conjunto de controle multivariado composto pelas cartas de Hotelling T^2 e de resíduos Q nos espaços reduzidos (KOURTI; MACGREGOR, 1995; MACGREGOR et al., 1994; YAN; YAO, 2015). No entanto, a viabilidade dos métodos de monitoramento multivariado baseados em MPCA é fortemente comprometida em situações nas quais o tamanho das JK variáveis desdobradas é equivalente ou maior que o tamanho das observações (ou bateladas) I, com $JK/I \rightarrow \infty$ (JOHNSTONE; LU, 2009; LEE; LEE; PARK, 2012). Isto ocorre porque, nesses casos, o PCA tradicional produz resultados inconsistentes, visto que a matriz de covariância amostral se torna um estimador notoriamente deficiente, com uma estrutura de autovalores e autovetores diferente da população original (AMINI, 2011; WANG; FAN, 2017). Assim, a busca por métodos que lidem adequadamente com a grande disponibilidade de dados fornecidos pelos processos computadorizados tem recebido destaque na literatura (CAPIZZI, 2015; WOODALL; MONTGOMERY, 2014). Um aumento significativo nos estudos desta área de conhecimento tem sido verificado, sendo a integração com métodos de seleção de variáveis um tópico em crescente desenvolvimento, principalmente nos últimos 5 anos (PERES; FOGLIATTO, 2018). Ainda assim, são escassas as aplicações identificadas na literatura quando o foco de interesse é o aprimoramento do monitoramento de processo em bateladas (CHU; QIN; HAN, 2004; YAN; KUANG; YAO, 2017; ZARZO; FERRER, 2004), o que justifica a necessidade de um maior aprofundamento deste tópico (PERES; FOGLIATTO, 2018).

Em relação ao objetivo principal desta tese, destaca-se a importância deste desenvolvimento, tanto como base para futuros desenvolvimentos acadêmicos quanto para a aplicação industrial destes novos métodos. O novo método '*Seleção de Variáveis de Pareto integrada a Análise de Componentes Principais Multidirecional'* (PVS-MPCA ou *Pareto Variable Selection – Multiway Principal Component Analysis*) almeja minimizar a emissão de alarmes falsos e, consequentemente, mitigar a limitação prática dos gráficos de contribuição, de recorrer a todas as variáveis originais para isolar as variáveis responsáveis pela falha detectada no processo, auxiliando no posterior diagnóstico e restauração da conformidade. Ao se analisar somente um número reduzido de variáveis, torna-se mais fácil e rápido identificar as responsáveis por um evento especial (WANG; JIANG, 2009), evitando-se a elevação dos custos do processo ou a venda de um produto de qualidade inferior ao usuário final (NOMIKOS; MACGREGOR, 1995a).

1.4 DELINEAMENTO DO ESTUDO

Definidos os objetivos da tese e apresentada a justificativa da importância desta pesquisa, esta seção estabelece o delineamento do estudo pelo qual esses objetivos serão alcançados, considerando o método de pesquisa e o método de trabalho utilizados.

1.4.1 Método de Pesquisa

De acordo com a forma de abordagem do problema, a pesquisa realizada nesta tese é classificada como quantitativa. Este tipo de abordagem baseia-se em métodos lógico-dedutivos que buscam explicar relações de causa/efeito e, através da generalização de resultados, possibilitar replicações (BERTO; NAKANO, 2000). O ato de mensurar variáveis de pesquisa é a característica mais marcante da abordagem quantitativa (MIGUEL et al., 2012).

O método científico aplicado na elaboração dos artigos é o hipotético-dedutivo, que se inicia pela percepção de uma lacuna nos conhecimentos, impossibilitando a explicação de um fenômeno e originando um problema de pesquisa. Para tentar solucionar esse problema são formuladas hipóteses, e evidências empíricas que invalidem a hipótese são buscadas. Quando não é possível demonstrar qualquer caso concreto capaz de derrubar a hipótese, tem-se a sua corroboração, a qual não excede o nível do provisório. Assim, a hipótese torna-se válida, pois superou todos os testes, mas não definitivamente confirmada, já que qualquer momento poderá surgir um fato que a invalide (GIL, 2008; MARCONI; LAKATOS, 2003).

Em relação aos objetivos, esta tese é classificada como pesquisa exploratória e aplicada. Segundo Gil (2008), a pesquisa exploratória tem como principal finalidade o esclarecimento e delimitação de um tema buscando desenvolver, elucidar e modificar conceitos e ideias a fim de proporcionar uma nova visão do problema. Dessa forma, é possível melhorar a compreensão do mesmo ou construir hipóteses pesquisáveis, passíveis de investigação mediante procedimentos mais sistematizados. A natureza aplicada se deve ao interesse na aplicação, utilização e consequências práticas dos conhecimentos gerados buscando solucionar problemas específicos, como as limitações encontradas no monitoramento de processos industriais em bateladas de elevada dimensionalidade.

1.4.2 Método de Trabalho

O desenvolvimento deste trabalho é realizado a partir de três artigos com objetivos específicos, os quais auxiliam o atingimento do objetivo geral da tese. Cada artigo e objetivo a ser alcançado faz uso de um método de trabalho específico. A estrutura do trabalho, os temas dos artigos, seus objetivos, questões de pesquisa e métodos são apresentados na Tabela 1.1.

Cabe ressaltar que os artigos são apresentados no formato de submissão aos periódicos internacionais estando, portanto, escritos em língua inglesa.

Tabela 1. 1 Estrutura das etapas da pesquisa desenvolvida

Estudos	Objetivos	Questões de Pesquisa	Revisão Teórica	Método de Pesquisa
Artigo 1 [(a)]	Identificar métodos que integram a seleção de variáveis ao controle estatístico de processo multivariado	1. Qual o *status* atual dos métodos que integram a seleção de variáveis com métodos de MSPC, e quais limitações os mesmos se propõem a superar?	1. Limitações dos métodos de MSPC 2. Métodos de Seleção de Variáveis 3. Etapas de monitoramento no Controle Estatístico de Processo	Pesquisa qualitativa: 1. Revisão sistemática de bibliografia
Artigo 2 [(b)]	Definir o método de alinhamento e sincronização de variáveis mais adequado para um banco de dados multivariado de um processo industrial em bateladas com duração variável	2. Qual o método de sincronização e alinhamento de dados de bateladas é mais adequado, visando ao aprimoramento da construção do modelo de monitoramento na Fase I?	1.Monitoramento de processos em bateladas 2. Alinhamento e sincronização de dados de bateladas através do DTW 3.Técnica de classificação por kNN 4.Processo de fabricação do chocolate	Pesquisa quantitativa: 1. Análise comparativa dos métodos propostos na literatura baseada nos resultados obtidos pela técnica de classificação por kNN
Artigo 3 [(c)]	Desenvolver um método que integre a seleção de variáveis às CCM baseadas em PCA visando a melhorar a detecção de falhas em bancos de dados de elevada dimensionalidade do processo industrial em bateladas	3. É possível melhorar o desempenho na detecção de falhas quando bancos de dados de processos em bateladas de elevada dimensionalidade são monitorados por um método que integra o MSPC coma a seleção de variáveis?	1. Métodos de seleção de variáveis com propósito de classificação de bateladas 2. CCM baseadas em PCA 3. Critérios de avaliação de desempenho de cartas de controle	Pesquisa quantitativa: 1. Comparar o desempenho do método proposto no monitoramento de um banco de dados de elevada dimensionalidade (obtido para o estudo de caso) com o monitoramento mediante o método tradicional usando CCM baseadas em PCA.

(a) Artigo publicado no periódico Computers & Industrial Engineering.

(b) Artigo submetido ao periódico Journal of Food Science and Technology, em fase de revisão.

(c) Artigo em fase de submissão.

O Artigo 1 - *Variable selection methods in multivariate statistical process control: a systematic literature review* (Métodos de seleção de variáveis no controle estatístico de processo multivariado: uma revisão sistemática de literatura) – busca, a partir de uma revisão sistemática de literatura, identificar: (*i*) as limitações existentes nos métodos de MSPC, (*ii*) os métodos de seleção de variáveis integrados ao MSPC para superar essas limitações, e (*iii*) as etapas do monitoramento estatístico de processo mais abordadas por esses métodos integrados. Mediante pesquisa e seleção de artigos relacionados ao tema, os métodos foram identificados, classificados e descritos. O artigo contribui ao apresentar a evolução do estado da arte do tema e inova ao propor (*i*) a classificação das metodologias de acordo com a abordagem de seleção de variáveis utilizada, e (*ii*) a categorização dos estudos de acordo com seu objetivo e etapa de monitoramento de processo para o qual foi desenvolvido. Assim, *clusters* de trabalhos foram propostos, auxiliando na identificação de lacunas e desdobramento de oportunidades de pesquisa sobre o tema.

O Artigo 2 – *Strategies for synchronizing chocolate conching batch process data using dynamic time warping* (Estratégias para sincronizar dados de processos em bateladas da conchagem do chocolate utilizando alinhamento temporal dinâmico) – busca selecionar o método de alinhamento e sincronização mais adequado para um banco de dados obtido de um processo de conchagem do chocolate, que apresenta bateladas com duração variável. O alinhamento e sincronização são necessários para que métodos de MSPC possam ser aplicados ao banco de dados. Em um banco de dados industrial do processo em bateladas da etapa de conchagem do chocolate ao leite foram aplicados três métodos baseadas no alinhamento temporal dinâmico (DTW ou *Dynamic Time Warping*), reportados por Kassidas, MacGregor e Taylor (1998), Ramaker et al. (2003) e González-Martínez, Ferrer e Westerhuis (2011). Os resultados são discutidos sob três pontos de vista: (*i*) da tecnologia de fabricação do chocolate, (*ii*) do poder de classificação das bateladas em conformes e não conformes mediante aplicação do método de classificação por *k*-vizinhos mais próximos (*kNN* ou *k nearest neighbors*), e (*iii*) do método mais adequado para tratamento de dados visando a aprimorar a construção do modelo de monitoramento na Fase I. Os três métodos se mostraram hábeis para promover o alinhamento e a sincronização, sendo o que apresentou os maiores valores das métricas de desempenho foi indicado como o mais adequado ao banco de dados analisado.

O Artigo 3 – *Fault detection in batch processes through variable selection integrated to multiway principal component analysis* (Detecção de falhas em processos em bateladas através da integração da seleção de variáveis à análise de componentes principais multidirecional) – propõe um método de detecção de falhas baseado nas

CCM T^2 e Q que lide com bancos de dados em bateladas de elevada dimensionalidade. Isso é alcançado mediante a aplicação do método de Seleção de Variáveis de Pareto (PVS ou *Pareto Variable Selection*), proposto por Anzanello et al. (2012), que seleciona um número reduzido de variáveis capazes de maximizar a acurácia de classificação das bateladas em conformes e não conformes. Posteriormente, esse subconjunto de variáveis selecionadas é utilizado na construção do modelo de referência para monitoramento das bateladas futuras. A escolha das variáveis selecionadas pelo método PVS foi corroborada pela análise técnica do processo de conchagem do chocolate ao leite, utilizado no estudo de caso. A melhora do desempenho da detecção de falhas obtida pela aplicação do método proposto, quando comparado ao método de CC baseadas em MPCA, demonstrou que as limitações do método tradicional foram superadas quando os bancos de dados com um número de variáveis muito superior ao de observações foi analisado.

1.5 DELIMITAÇÕES DO ESTUDO

O presente trabalho se concentra na análise de lacunas relacionadas a três temas relevantes para o controle de processos.

Na área de MSPC, o foco está na melhoria da detecção de falhas pelas CCM T^2 e Q baseadas em MPCA. O monitoramento através de CCM de soma acumulada (MCUSUM) e média móvel exponencialmente ponderada (MEWMA) (BERSIMIS; PSARAKIS; PANARETOS, 2007) não fazem parte do escopo desta tese. Também estão excluídas as etapas de isolamento de variáveis, diagnóstico de falhas e intervenção no processo industrial, compreendidas na Fase II do monitoramento (WOODALL; MONTGOMERY, 2014).

No que tange a seleção de variáveis, somente métodos com fins de classificação serão avaliados (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012), não sendo considerados métodos com propósito preditivo. A implementação da seleção de variáveis se dará mediante uso da abordagem *wrapper*, na qual o subconjunto de variáveis relevantes será determinado através de um procedimento iterativo envolvendo um *ranking* de importância de variável e um algoritmo para classificação de bateladas. Abordagens de filtro pré- e pós-processamento, e embarcada (GHOSH; RAMTEKE; SRINIVASAN, 2014; MEHMOOD et al., 2012) não serão abordadas.

Por fim, o método proposto é delimitado para monitoramento *off-line* (KOURTI, 2003) e para processos industriais em bateladas. Um conceito mais amplo, que avalie processos contínuos em tempo real, não se encontra no escopo deste trabalho devendo ser alvo de pesquisas futuras.

1.6 ESTRUTURA DA TESE

Esta tese está organizada em cinco capítulos principais. No Capítulo 1 foram introduzidos o problema, o tema a ser desenvolvido, bem como os objetivos, justificando a importância da pesquisa dos pontos de vista acadêmico e prático. O capítulo também apresentou o método de trabalho, a estrutura e as delimitações do estudo. Na sequência, os capítulos 2, 3 e 4 apresentam os artigos desenvolvidos, conforme estrutura apresentada na Tabela 1.1. O Capítulo 5 aborda as conclusões da tese e sugestões de pesquisas futuras a serem desenvolvidas a partir dos resultados apresentados.

1.7 REFERÊNCIAS

AMINI, A. A. **High-dimensional principal component analysis**. 2011. 163p. Tese (Doutorado) - Electrical Engineering and Computer Sciences, University of California, Berkley.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, p. 97–105, 2012.

ANZANELLO, M. J.; FOGLIATTO, F. S. A review of recent variable selection methods in industrial and chemometrics applications. **European Journal of Industrial Engineering**, v. 8, n. 5, p. 619–645, 2014.

BERSIMIS, S.; PSARAKIS, S.; PANARETOS, J. Multivariate statistical process control charts: An overview. **Quality and Reliability Engineering International**, v. 23, p. 517–543, 2007.

BERTO, R. M. V. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Produção**, v. 9, n. 2, p. 65–75, 2000.

BISGAARD, S. The future of quality technology: From a manufacturing to a knowledge economy & from defects to innovations. **Quality Engineering**, v. 24, p. 30–36, 2012.

CAPIZZI, G. Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II. **Quality Engineering**, v. 27, p. 44-67, 2015.

CHIANG, L. H.; KOTANCHEK, M. E.; KORDON, A. K. Fault diagnosis based on Fisher discriminant analysis and support vector machines. **Computers and Chemical Engineering**, v. 28, p. 1389–1401, 2004.

CHU, Y.-H.; QIN, S. J.; HAN, C. Fault Detection and Operation Mode Identification Based on Pattern Classification with Variable Selection. **Industrial & Engineering Chemistry Research**, v. 43, p. 1701–1710, 2004.

GARCÍA-MUÑOZ, S.; KOURTI, T.; MACGREGOR, J. F.; MATEOS, A. G.; MURPHY, G. Troubleshooting of an Industrial Batch Process Using Multivariate Methods. **Industrial and Engineering Chemistry Research**, v. 42, p. 3592–3601, 2003.

GHOSH, K.; RAMTEKE, M.; SRINIVASAN, R. Optimal variable selection for effective statistical process monitoring. **Computers and Chemical Engineering**, v. 60, p. 260–276, 2014.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6ed. São Paulo: Atlas S.A., 2008. 200p.

GONZÁLEZ-MARTÍNEZ, J. M.; FERRER, A.; WESTERHUIS, J. A. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. **Chemometrics and Intelligent Laboratory Systems**, v. 105, p. 195–206, 2011.

JIANG, Q.; YAN, X.; HUANG, B. Performance-Driven Distributed PCA Process Monitoring Based on Fault-Relevant Variable Selection and Bayesian Inference. **IEEE Transactions on Industrial Electronics**, v. 63, n. 1, p. 377–386, 2016.

JIANG, W.; WANG, K.; TSUNG, F. A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. **Journal of Quality Technology**, v. 44, n. 3, p. 209–230, 2012.

JOHNSTONE, I. M.; LU, A. Y. On consistency and sparsity for principal components analysis in high dimensions. **Journal of the American Statistical Association**, v. 104, n. 486, p. 682–693, 2009.

KASSIDAS, A.; MACGREGOR, J. F.; TAYLOR, P. A. Synchronization of batch trajectories using dynamic time warping. **AIChE Journal**, v. 44, n. 4, p. 864–875, 1998.

KOURTI, T. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. **Annual Reviews in Control**, v. 27, p. 131–139, 2003.

KOURTI, T.; MACGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics and Intelligent Laboratory Systems**, v. 28, p. 3–21, 1995.

KOURTI, T.; NOMIKOS, P.; MACGREGOR, J. F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. **Journal of Process Control**, v. 5, n. 4, p. 277–284, 1995.

KUANG, T. H.; YAN, Z.; YAO, Y. Multivariate fault isolation via variable selection in discriminant analysis. **Journal of Process Control**, v. 35, p. 30–40, 2015.

LEE, Y. K.; LEE, E. R.; PARK, B. U. Principal component analysis in very high-dimensional spaces. **Statistica Sinica**, v. 22, p. 933–956, 2012.

MACGREGOR, J. F.; JAECKLE, C.; KIPARISSIDES, C.; KOUTOUDI, M. Process monitoring and diagnosis by multiblock PLS methods. **AIChE Journal**, v. 40, n. 5, p. 826–838, 1994.

MARCONI, M.; LAKATOS, E. **Fundamentos de metodologia científica**. 5ed. São Paulo: Atlas S.A., 2003. 311p.

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, v. 23, p. 35–44, 1999.

MEGAHED, F. M.; JONES-FARMER, A. A statistical process monitoring perspective on “big data”. In **Frontiers in Statistical Quality Control**. 11th ed., 2013. 21p.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SæBØ, S. A review of variable selection methods in Partial Least Squares Regression. **Chemometrics and Intelligent Laboratory Systems**, v. 118, p. 62–69, 2012.

MIGUEL, P.A.C.; FLEURY, A.; MELLO, C.H.P.; NAKANO, D. N.; LIMA, E.P.; TURRIONI, J.B.; HO, L.L.; MORABITO, R.; MARTINS, R.A.; SOUSA, R.; COSTA, S.E.G.; PUREZA, V. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 2ed. Rio de Janeiro: Elsevier, 2012. 260p.

NOMIKOS, P.; MACGREGOR, J. F. Monitoring batch processes using multiway principal component analysis. **AlChE Journal**, v. 40, n. 8, p. 1361–1375, 1994.

NOMIKOS, P.; MACGREGOR, J. F. Multivariate statistical process control charts for monitoring batch processes. **Technometrics**, v.37, n. 1, p. 41–59, 1995a.

NOMIKOS, P.; MACGREGOR, J. F. Multi-way partial least squares in monitoring batch processes. **Chemometrics and Intelligent Laboratory Systems**, v. 30, p. 97–108, 1995b.

PERES, F. A. P.; FOGLIATTO, F. S. Variable selection methods in multivariate statistical process control: A systematic literature review. **Computers and Industrial Engineering**, v. 115, p. 603–619, 2018.

RAMAKER, H. J.; VAN SPRANG, E. N. M.; WESTERHUIS, J. A.; SMILDE, A. K. Dynamic time warping of spectroscopic batch data. **Analytica Chimica Acta**, v. 498, p. 133–153, 2003.

VAN SPRANG, E. N. M.; RAMAKER, H-J; WESTERHUIS, J. A.; GURDEN, S. P.; SMILDE, A. K. Critical evaluation of approaches for on-line batch process monitoring. **Chemical Engineering Science**, v. 57, p. 3979–3991, 2002.

WANG, K.; JIANG, W. High-dimensional process monitoring and fault isolation via variable selection. **Journal of Quality Technology**, v. 41, n. 3, p. 247–258, 2009.

WANG, W.; FAN, J. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. **The Annals of Mathematical Statitsics**, v. 45, n. 3, p. 1342–1374, 2017.

WOODALL, W.; MONTGOMERY, D. Some current directions in the theory and application of statistical process monitoring. **Journal of Quality Technology**, v. 46, n. 1, p. 78–94, 2014.

YAN, Z.; KUANG, T. H.; YAO, Y. Multivariate fault isolation of batch processes via variable selection in partial least squares discriminant analysis. **ISA Transactions**, v. 70, p. 389–399, 2017.

YAN, Z.; YAO, Y. Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). **Chemometrics and Intelligent Laboratory Systems**, v. 146, p. 136–146, 2015.

ZARZO, M.; FERRER, A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. **Chemometrics and Intelligent Laboratory Systems**, v. 73, p. 15–27, 2004.

2 ARTIGO 1 – VARIABLE SELECTION METHODS IN MULTIVARIATE STATISTICAL PROCESS CONTROL: A SYSTEMATIC LITERATURE REVIEW

Artigo publicado em Computers & Industrial Engineering 115 (2018)

Abstract

Technological advances led to increasingly larger industrial quality-related datasets calling for process monitoring methods able to handle them. In such context, the application of variable selection (VS) in quality control methods emerges as a promising research topic. This review aims at presenting the current state-of-the-art of the integration of VS in multivariate statistical process control (MSPC) methods. Proposals aligned with the objective were identified, classified according to VS approach, and briefly presented. Research on the topic has considerably increased in the past five years. Thirty methods were identified and categorized in 10 clusters, according to the objective of improvement in MSPC and the step of process monitoring they were aimed to improve. The majority of the propositions were either targeted at exclusively monitoring potential out-of-control variables or improving the monitoring of in-control variables. MSPC improvements were centered in principal component analysis (PCA) projection methods, while VS was mainly carried out using the Least Absolute Shrinkage and Selection Operator (LASSO) method and genetic algorithms. Fault isolation was the most addressed step in process monitoring. We close the paper proposing five topics for future research, exploring the opportunities identified in the literature.

Keywords: Variable selection. Multivariate statistical process control. Industrial process monitoring. High dimensional dataset.

2.1 INTRODUCTION

In recent decades, technological advances significantly reduced costs and barriers related to information collection and storage in industrial environments. Consequently, databases with readings from hundreds or thousands of variables describing the behavior of industrial processes have become available, calling for the development of new multivariate statistical process control (MSPC) methods (MEGAHED; JONES-FARMER, 2013; MEHMOOD et al., 2012; VAN AELST; WELSCH; ZAMAR, 2010). Traditionally, multivariate control charts (MCCs) based on projection methods such as Principal Component Analysis (PCA) or Partial Least Squares (PLS) regression have been used to monitor multivariate processes. In those charts, after an out-of-control (OOC) alarm is triggered, the projected point is decomposed in its original variables, which are then analyzed using Contribution Plots to determine which variables are responsible for the alarm. As the dimensionality of the database under analysis increases, the decomposition step becomes infeasible due to the extensive work involved in the construction and interpretation of Contribution Plots. In such scenarios, the integration of VS methods to MSPC approaches become a promising research topic (KOURTI, 2005; MARTIN; MORRIS; KIPARISSIDES, 1999; MEGAHED; JONES-FARMER, 2013; MEHMOOD et al., 2012).

The main objective of VS methods is to identify a subset of variables that carries most of the relevant information contained in the complete dataset. Some common VS methods applied in industrial datasets are forward selection (FS) and backward selection techniques, data mining tools, PLS, PCA, and clustering. Optimization tools, such as linear programming and genetic algorithms (GA), have also found wide application in the analysis of more complex systems (ANZANELLO; FOGLIATTO, 2014).

The improvement of MCCs through integration with VS methods has been discussed by Capizzi (2015), who reviewed the efficiency and advantages of the combined approach when monitoring only a subset of variables that are potentially responsible for a fault alarm. However, the scope of methods integrating VS and MSPC is much broader, including not only MCCs to monitor OOC variables, but also several frameworks to promote the improvement of process monitoring. One other review by Anzanello and Fogliatto (2014) covered relevant VS methods in Chemometrics and industrial applications, aiming at a better prediction of continuous and categorical response variables; their review, however, did not cover works that propose VS as a means to attain MSPC improvement.

This paper is the first to present the current state-of-the-art on VS methods integrated to MSPC through a systematic review. We provide answers to the following research questions: (*i*) which limitations in MSPC methods should be overcome?, (*ii*) which VS methods are used to improve MSPC?, (*iii*) which steps of process monitoring in statistical process control (SPC) were studied?, and (*iv*) which research opportunities arise from gaps in the current state-of-the-art on the subject?. To answer those questions, methods available in the literature were identified, grouped according to similarity, and presented. It is not our objective to explain in depth the mathematical fundamentals of methods revised, but to provide a sufficient description that allows their comparison and visualization of deployments proposed.

This article is organized in five sections, in addition to the present introduction. In Section 2.2, the methodology used for the systematic review is presented. Study characterization is given in section 2.3. The proposed VS-MSPC integration methods are presented in section 2.4, and process monitoring in SPC and performance of developed methods in section 2.5. Finally, conclusions and research opportunities are given in section 2.6. Table 2.1 shows the acronyms used in this article.

Table 2. 1 Acronyms used in the article

Acronym	Description	Acronym	Description
AIC	Akaike information criterion	MEWMA	Multivariate exponentially weighted moving average
ADR	Adaptive dimension reduction	MKPCA	multi-model kernel PCA
ARL	Average run length	MPLS	Multiway partial least squares
BBGVS	Bootstrapping-based generalized variable selection	MRR	Missing reconstruction ratio
BSPCA	Bayesian subspace PCA	MSPC	Multivariate statistical process control
CC	Control chart	MSN	Multivariate standardized shift
CI	Combined index	NFDI	Nonlinear fault detection index
CUSUM	Cumulative sum	NIR	Near infrared
2-D-DPCA	Two-dimensional dynamic principal component analysis	NSGA-II-JG	Non-dominated sorting genetic algorithm and a jumping gene operator
DISSIM	Dissimilarity	OOC	Out-of-control
EN	Elastic net	OPA	Orthogonal projection approach
EWMA	Exponentially weighted moving average	PCA	Principal component analysis
FA	Factor analysis	PCR	Principal component regression
FAR	False alarm rate	PCS	Principal component subspace
FBPCA	Fault-bayesian PCA	PLS	Partial least squares
FDA	Fisher’s discriminant analysis	RMSEP	Root-mean-square error in prediction
FIR	Fuzzy inductive reasoning	ROS	Region of Support
FS	Forward selection	RS	Residual subspace
GA	Genetic algorithm	SFFS	Sequential forward floating selection
GLR	Generalized log-likelihood ratio	SDISSIM	Sparse dissimilarity
IC	In-control	SPC	Statistical process control
LAR	Least angle regression	SPLS	Sparse partial least squares
LARSEN	Least angle regression and elastic net algorithm	SR	Spatial rank

Acronym (*continue*)	Description	Acronym (*continue*)	Description
LASSO	Least absolute shrinkage and selection operator	SVM	Support vector machine
LEWMA	LASSO-based EWMA	TDB	Two-dimensional Bayesian
MCC	Multivariate control chart	TEP	Tennessee eastman process
MCUSUM	Multivariate cumulative sum	U-PLS	Unfold PLS
MDR	Missed detection rate	VS	Variable selection

2.2 METHOD

The aim of this article is to systematically review the literature on VS methods integrated to MSPC, guided by the research questions in section 2.1. To select the group of articles to be covered in this review, a series of steps was adopted to ensure appropriate rigor and repeatability.

Databases surveyed were Science Direct and Web of Science. The choice was restricted to these two databases since they host all relevant JCR-indexed journals in the field of quality control. Articles in English were considered. Keywords used in the search were: (“*variable selection*”) AND ((*multivariate "statistical process control"*) OR (*"fault monitoring"*) OR (*“monitoring process”*) OR (*“process monitoring”*) OR (*“monitoring system”*)) OR ((*batch*)) OR ((*"manufacturing applications"*) OR (*"industrial applications"*) OR (*“discrete manufacturing”*)), provided it was present in article’s title, abstract or keywords. Boolean operators “*AND*” and “*OR*” were used to allow combining groups of words in the search. Only articles published in scientific journals were considered. Search in databases took place on September 28, 2017; no restriction was imposed on publication timespan. Exclusion criteria were: (*i*) repeated articles, and (*ii*) articles that did not mention the integration of VS methods and MSPC in the title or abstract. The final group of articles was entirely read, such that results could be presented and discussed.

The sequence of steps described above, and the number of items found in each step are given in Figure 2.1.

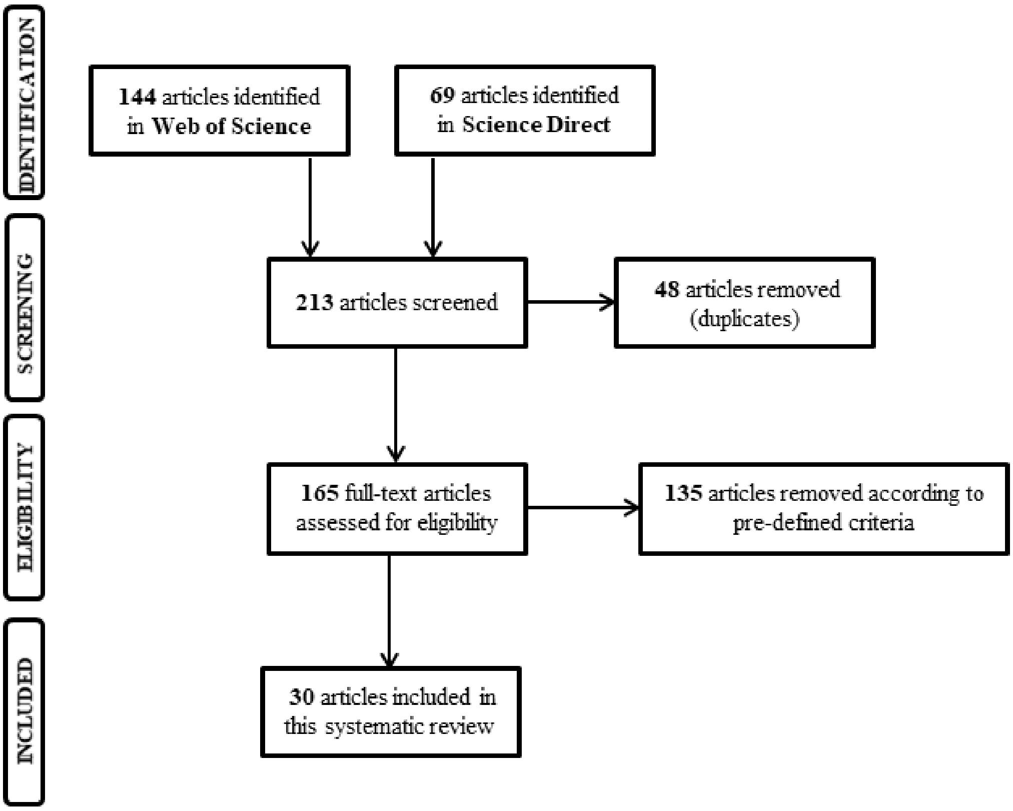

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
144 articles identified in Web of Science
69 articles identified in Science Direct
213 articles screened
48 articles removed (duplicates)
165 full-text articles assessed for eligibility
135 articles removed according to pre-defined criteria
30 articles included in this systematic review

Figure 2. 1 Search steps and results

2.3 STUDY CHARACTERIZATION

The number of studies addressing the integration of MSPC and VS methods revealed that the topic has been the subject of a growing number of articles in recent years. Table 2.2 presents the articles included in this review ordered by year of publication, and divided in three time periods, displaying an important increase in the number of publications in the most recent period. Articles are also identified according to journal title and country of origin.

Table 2. 2 Evolution by year, journal title and country of origin of selected articles

Authors	Year	Journal Title	Country	Number of Articles per Period		
				2000-2005	2006-2011	2012-2017
Tur et al.	2002	International Journal of General Systems	Spain and USA	**6**		
Chu, Lee and Han	2004	Industrial & Engineering Chemistry Research	South Korea			
Gourvénec, Capron, and Massart	2004	Analytica Chimica Acta	Belgium			
Zarzo and Ferrer	2004	Chemometrics and Intelligent Laboratory Systems	Spain			
Chiang, Pell, and Seasholtz	2004	IFAC Proceedings Volumes	USA			
Chu, Qin and Han	2004	Industrial & Engineering Chemistry Research	Korea and USA			
Wang and Jiang	2009	Journal of Quality Technology	China		**8**	
Yao et al.	2009	Industrial & Engineering Chemistry Research	China			
Zou and Qiu	2009	Journal of the American Statistical Association	China and USA			
Wang and Tsung	2009	Quality and Reliability Engineering International	China			
González and Sánchez	2010	Journal of Quality Technology	Spain			
Ge, Zhang, and Song	2010	Journal of Process Control	China			
Capizzi and Masarotto	2011	Technometrics	Italy			
Ge, Gao, and Song	2011	Chemical Engineering Science	China			
Jiang, Wang, and Tsung	2012	Journal of Quality Technology	China			**16**
Jeong et al.	2012	International Journal of Hydrogen Energy	South Korea			
Zou, Ning, and Tsung	2012	Annals of Operations Research	China			
Ghosh, Ramteke, and Srinivasan	2014	Computers and Chemical Engineering	Singapore			
Giannetti et al.	2014	Computers & Industrial Engineering	United Kingdom			
Yan and Yao	2015	Chemometrics and Intelligent Laboratory Systems	China and Taiwan			
Nishimura, Matsuura, and Suzuki	2015	Statistics & Probability Letters	Japan			
Kuang, Yan, and Yao	2015	Journal of Process Control	China and Taiwan			
Jiang, Yan, and Huang	2016	IEEE Transactions on Industrial Electronics	China and Canada			
Zhao and Wang	2016	Journal of Process Control	China			
Jiang and Huang	2016	Journal of Process Control	Canada			
Li et al.	2017	Computers & Industrial Engineering	China			
Abdella et al.	2017	Quality and Reliability Engineering International	Qatar and USA			
Shinozaki and Iida	2017	Communications in Statistics – Theory and Methods	Japan			
Yan, Kuang, and Yao	2017	ISA Transactions	China and Taiwan			
Zhao and Gao	2017	Control Engineering Practice	China			

In the first period (2000 – 2005), most of the methods were applied to batch industrial processes, and the VS was proposed with the aim of creating new frameworks to improve modeling and prediction tasks performed using Fuzzy Inductive Reasoning (FIR) and PLS regression, and to improve monitoring of normal observations using the Orthogonal Projection Approach (OPA), and classification methods, such as Fisher Discriminant Analysis (FDA) and Support Vector Machine (SVM). The enhancement of MCCs and the monitoring of processes through PCA were introduced in the second period (2006 – 2011), and investigated in further depth in the third period (2012 – 2017), along with articles about PLS regression, the dissimilarity (DISSIM) method, and the T^2 test. The evolution in the number papers published in these three periods corroborates the growing interest on the subject in the literature. From 2000 to 2005, there were 6 published papers (averaging 1.0 per year); from 2006 to 2011 the average increased to 1.3 papers/year, totaling 8 papers; in the most recent period (2012-2017) the number of papers doubled, averaging 2.7 papers/year, and accounting for 53.3% of the articles selected for this review.

China was the country of origin of most authors, contributing with 50% of the papers, followed by the United States of America, with 5 papers, Spain and Taiwan with 3 papers each, and South Korea and Japan, with 2 papers each. Articles reviewed here were published in 20 different journals. The largest number of articles appeared in *Journal of Process Control* (4), followed by *Journal of Quality Technology* (3), and *Computers & Industrial Engineering*, *Chemometrics and Intelligent Laboratory Systems*, *Industrial & Engineering Chemistry Research*, and *Quality and Reliability Engineering International* (2 each). Remaining journals presented one publication each, displaying a diversity of applications on the subject.

Limitations encountered in the application of MSPC methods and specific research objectives motivated by them are summarized in Table 2.3, providing an answer to our first research question (*'which limitations in MSPC methods should be overcome?'*).

Table 2. 3 Limitations of MSPC methods and specific objectives of articles covered in this review

Authors	Limitations	Objectives
Tur et al. (2002)	The Fuzzy Inductive Reasoning (FIR) technique is well suited for qualitative behavioral modeling and simulation of physical systems. However, due to its computational complexity, VS algorithms are required	Find a VS algorithm with lower computational complexity, in order to select a set of candidate input variables and reduce the model search space of FIR yielding high predictability and specificity qualitative models for the system outputs
Chu, Lee, and Han (2004)	The use of all unfolded variables from a three-way batch process data matrix may impair the prediction performance of PLS models, since only a small portion of those process variables are correlated to quality response variables	Improve the prediction performance of PLS models in batch processes through the selection of process variables related to quality response variables
Gourvénec, Capron, and Massart (2004)	In chemical process monitoring, the time required to acquire one spectrum and to transfer it to the database is too high. Moreover, the number of available spectra during a certain period could decrease if many absorbances at several wavelengths are acquired	Implement VS to yield smaller spectra, which allows the acquisition and analysis of a higher volume of data in a given period, and improves the online prediction of concentration profiles
Zarzo and Ferrer (2004)	In a batch process, when several process variables and the quality output are affected by a fault, all variables are considered to be correlated with the quality response variable, but it is difficult to reveal which process variables are actually responsible for the variation in process quality	Select variables that potentially affect the quality of a chemical batch process, allowing the determination of optimal process conditions using Design of Experiments
Chiang, Pell, and Seasholtz (2004)	The contribution charts perform well in simple faults' identification, but are less effective in identifying complex process faults	Develop an alternative method for identifying process faults
Chu, Qin, and Han (2004)	Due to the multimodal distribution of normal data in processes with multiple operation modes, fault detection using PCA may be seriously compromised	Propose a novel method for improved fault detection in multimode operation along with the proper identification of operation modes
Wang and Jiang (2009)	In high dimensional datasets it is difficult to detect potential process shifts and locate root causes of faults	Propose the VS-MSPC chart to monitor variables that are probably responsible for OOC alarms, simultaneously improving fault detection performance and isolating their root causes
Yao et al. (2009)	The assumption that the support region (ROS) in a two-dimensional dynamic principal component analysis (2-D-DPCA) model is limited to the quarter plane and have a regular shape is not always reasonable in certain batch processes	Present a solution to the problem of ROS determination for the 2-D-DPCA model

Authors	Limitations (*continue*)	Objectives (*continue*)
Zou and Qiu (2009)	MSPC CC with quadratic charting statistics are powerful in detecting shifts occurring due to changes in the majority of components in multivariate process mean vector. However, in practice shifts are often due to only a few of them. Also, in high dimensional processes such control charts, as well as conventional fault isolation approaches, are infeasible, because the total number of possible shifts directions increases exponentially	Propose an MSPC control chart capable of detecting shifts in one or more mean vector components efficiently, implying in reasonably small computations and providing an effective post signal diagnostic tool for identifying shifted components
Wang and Tsung (2009)	The identification of non-informative variables is difficult in processes with dynamic shifts, as the contribution of each variable changes continuously over time	Propose an adaptive dimension reduction scheme that adjusts the dimensions of an MCC online based on real-time information collected from the process
González and Sánchez (2010)	When redundant variables are monitored, process control costs unnecessarily increase. It may also occur that a variable with large measurement error is measured instead of another highly correlated variable with lower measurement error	Select a subset of variables carrying the largest amount of information about the process, improving its monitoring and avoiding the increase in costs
Ge, Zhang, and Song (2010)	Traditional PCA-based monitoring methods assume that process variables are linear, normally distributed, and operated in single mode. In reality, those restrictions are easily violated in data obtained from complex processes	Develop an improved nonlinear process monitoring method
Capizzi and Masarotto (2011)	VS-MSPC and LEWMA CC considered cases in which shifts directly involve components of the mean vector, but not shifts occurring at some stage of a multistage process or shifts involving the regression coefficients of a general linear profile. Those CCs were also not able to detect increases in process variability	Develop a new CC for fault detection of shifts in both the mean and the total variability of multidimensional processes
Ge, Gao, and Song (2011)	Processes with nonlinear and multimode behaviors are usually monitored separately, calling for the development of a monitoring approach which considers both behaviors	Develop an efficient monitoring method for processes with both nonlinear and multimode characteristics
Jiang, Wang, and Tsung (2012)	VS-MSPC chart is a Shewhart-type chart that only uses information from the current process observation	Propose the VS-MEWMA chart that combines a VS strategy and the accumulation of recent process information
Jeong et al. (2012)	In an electricity generation system, false alarms occurs quite frequently and simultaneously, and faults involving drift of multiple variables are usually not detected	Reduce the number of false alarms and improve fault detection using a VS heuristic method based on PCA and Factor analysis

Authors	Limitations (*continue*)	Objectives (*continue*)
Zou, Ning, and Tsung (2012)	A drawback in the existing parametric profile monitoring methods is that if number of profile parameters is large, their detection ability tends to decline substantially. Moreover, it is also challenging to identify which parameter(s) have changed after an alarm is triggered	Design an efficient SPC scheme for multivariate profile monitoring and diagnosis
Ghosh, Ramteke, and Srinivasan (2014)	The use of all measured variables to build a monitoring model could impair process monitoring performance by MSPC methods	Built a reduced PCA model based on the subset of most relevant variables identified by GA, to maximize the monitoring performance in a multi-fault analysis
Giannetti et al. (2014)	Discovering the root causes of defects in the context of foundries by simultaneously analyzing process data containing a mixture of categorical and continuous variables is a challenging task	Extend the approach based on the use of co-linearity index and penalty matrix (RANSING et al., 2013) to data containing a mixture of continuous and categorical variables, and discover the optimal process settings that are most correlated with responses, improving fault diagnosis via PCA
Yan and Yao (2015)	In a process with massive amount of variables, when fault directions are unavailable and historical fault data are insufficient the computational burden to solve conventional reconstruction-based methods is often too heavy to be carried out online	Develop a method based on the LASSO algorithm to reconstruct variables that are potential responsible for faults, improving fault isolation via PCA
Nishimura, Matsuura, and Suzuki (2015)	VS-MSPC and the VS-MEWMA CC may have their performance impaired if the number of selected variables in VS is not equal, or nearly equal, to a predetermined number	Propose a new criterion to establish the number of selected variables in VS-MEWMA charts, resulting in the AIC-MEWMA CC
Kuang, Yan, and Yao (2015)	In traditional fault isolation methods faulty variables may influence the contributions of non-faulty variables (contribution plots), all candidate fault directions are assumed to be known (reconstruction-based methods), and sufficient historical fault data are required for model training (pattern classification techniques)	Promote another point of view for root-cause diagnosis through the development of a fault isolation method that provides information on the relevance of process variables for the detected faults
Jiang, Yan, and Huang (2016)	When using a single PCA model for all faults (GHOSH; RAMTEKE; SRINIVASAN, 2014) it is disregarded that the subset of relevant variables for one fault may yield poor monitoring performance of other faults	Develop a method for fault isolation based on the selection of optimal subsets of variables, allowing the modelling of fault effects
Zhao and Wang (2016)	Difficulty in deciding which process variables to include in a reconstruction model to more effectively explore fault effects, and thus correct the alarm signals for fault isolation	Apply the faulty VS idea to reconstruction modeling building a new method for fault isolation

Authors	Limitations (*continue*)	Objectives (*continue*)
Jiang and Huang (2016)	Traditional distributed monitoring schemes are employed to deal with large-scale processes that are assumed to have been properly decomposed. However, that is a difficult task and may not be true in several practical applications. Correct fault diagnosis in distributed monitoring is also an issue	Introduce a performance-driven process decomposition and a fault isolation system for distributed process monitoring
Li et al. (2017)	The VS-MSPC and VS-MEWMA requires known IC mean μ_0 and covariance matrix $\mathbf{\Sigma}$, or a large enough IC dataset to estimate them. Also in the modified VS-MEWMA the control limit cannot be obtained by simulation when the parameters of the underlying distribution are unknown	Propose the SR-VSMEWMA which explores a self-starting technique to substantially reduce the amount of IC data required to construct the VS-MEWMA CC
Abdella et al. (2017)	New CCs, such as VS-MEWMA, were developed to deal with the performance degradation of MCCs in high-dimensional SPC applications. However, the VS-MEWMA chart may deteriorate its performance in detecting small process changes when the VS procedure malfunctions	Develop a MCUSUM-based method to improve sensitivity to detect changes in the mean of process variables, improve the detection of small mean changes in the mean vector of multivariate normal processes, and provide useful information to identify faulty variables in high-dimensional processes
Shinozaki and Iida (2017)	Increasing the number of variables does not necessarily lead to increased power, or the probability that an abnormal item is detected, even when parameters of the distribution are known. Testing procedures proposed for increased power require VS methods to seek for the variables' subset with largest power	Handle the problem of detecting abnormal items based on a T^2 test, and propose a simple and effective VS method based on unbiased estimators of the detection power of subsets
Yan, Kuang, and Yao (2017)	LASSO-based and EN-based methods overcome drawbacks of contribution plots and reconstruction-based methods in fault isolation, but are not applicable to batch process data	Develop a multivariate fault isolation method that is particularly useful for batch process data analysis
Zhao and Gao (2017)	The DISSIM method has been successfully used for detection of incipient faults, but fault isolation of abnormal variables that distort the variable covariance structure has not been well addressed	Develop a variable isolation procedure that takes into account the data distribution structure and does not need any *a priori* fault knowledge

Fifteen MSPC and analytical methods were adapted to overcome the limitations of monitoring high dimensional industrial datasets. Most improvements (13 of 30 methods) targeted at PCA and PLS projection methods. The improvement in exponentially weighted moving average (EWMA)-based methods was reported in five papers; reconstruction-based methods and FDA were addressed by two methods each.

Evaluating the objectives reported, three main goals in adaptations of MSPC methods were identified; they were: exclusive monitoring of potential OOC variables (ABDELLA et al., 2017; CAPIZZI; MASAROTTO, 2011; JIANG; WANG; TSUNG, 2012; KUANG; YAN; YAO, 2015; LI et al., 2017; NISHIMURA; MATSUURA; SUZUKI, 2015; SHINOZAKI; IIDA, 2017; WANG; JIANG, 2009; YAN; KUANG; YAO, 2017; YAN; YAO, 2015; ZHAO; WANG, 2016; ZOU; NING; TSUNG, 2012; ZOU; QIU, 2009) better modeling and prediction of response variables (CHU; LEE; HAN, 2004; TUR et al., 2002; ZARZO; FERRER, 2004), and improvement in the monitoring of in-control (IC) variables (CHIANG; PELL; SEASHOLTZ, 2004; CHU; QIN; HAN, 2004; GE; GAO; SONG, 2011; GE; ZHANG; SONG, 2010; GHOSH; RAMTEKE; SRINIVASAN, 2014; GIANNETTI et al., 2014; GONZÁLEZ; SÁNCHEZ, 2010; GOURVÉNEC; CAPRON; MASSART, 2004; JEONG et al., 2012; JIANG; HUANG, 2016; JIANG; YAN; HUANG, 2016; WANG; TSUNG, 2009; YAO et al., 2009; ZHAO; GAO, 2017).

2.4 PROPOSED VS-MSPC INTEGRATION METHODS

In this section, 30 methods identified in our search are classified and briefly presented. Classification was carried out according to the approach proposed to integrate VS into MSPC, namely: Filter and Wrapper. Filter-based approaches were divided according to their strategy regarding the VS step, as follows: Preprocessing, in which the complete set of variables was reduced to a subset of relevant ones prior to the application of the MSPC method, and Postprocessing, in which the subset of relevant variables was defined from the outputs of the monitoring models. In wrapper approaches the subset of relevant variables was determined through an iterative procedure involving the VS step and the MSPC method chosen for monitoring (GHOSH; RAMTEKE; SRINIVASAN, 2014; MEHMOOD et al., 2012). Among the 30 methods reviewed here, 16 were classified in the Preprocessing Filter class, 2 in the Postprocessing Filter class, and 12 in the Wrapper class. Figure 2.2 displays how VS and MSPC interact in each class.

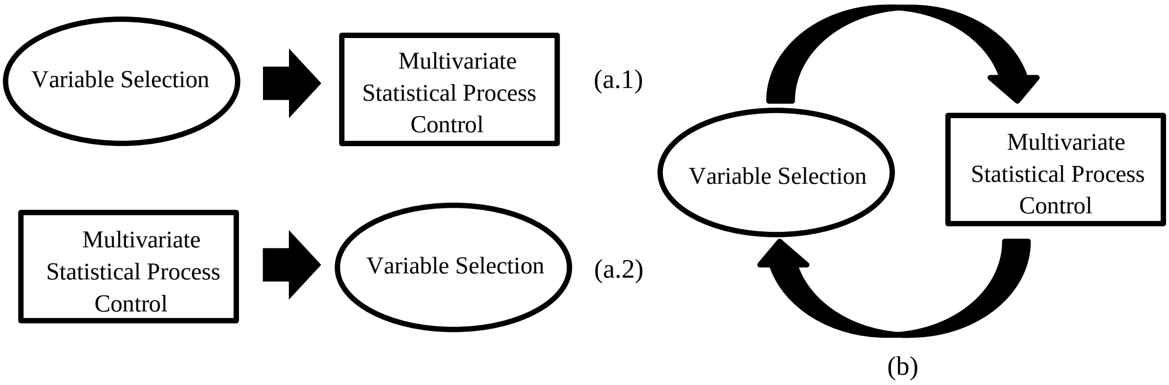

Variable Selection
Multivariate Statistical Process Control
(a.1)
Multivariate Statistical Process Control
Variable Selection
(a.2)
Variable Selection
Multivariate Statistical Process Control
(b)

Figure 2. 2 Interaction between VS and MSPC steps in class (a.1) Preprocessing Filter Approach, (a.2) Postprocessing Filter Approach, and (b) Wrapper Approach

A summary of the proposed methods informing the MSPC and VS strategies adopted in each case, in addition to the VS approach class, type of process, branch of industrial application, steps in SPC monitoring, and structure of the method they are aimed at is presented in Table 2.4. With that, our second research question (*'which VS methods are used to improve MSPC?'*) is addressed.

The most frequently used methods were Least Absolute Shrinkage and Selection Operator (LASSO) and GA, with five methods each. LASSO-based methods were developed to improve fault isolation through the development of two new EWMA-based control charts, one new reconstruction-based framework, one framework to improve the dissimilarity distribution concept, and one framework that uses discriminant analysis. With the objective of improving the monitoring of IC variables using PCA and FDA, GAs were applied to develop new frameworks able to deal with fault detection and isolation. FS was proposed in four studies to improve the performance of MCCs through the selection of OOC variables that should be monitored. Remaining methods were based on different VS methods, which will be presented in subsections to follow.

Table 2. 4 Main characteristics of methods reviewed

		Proposed Method	Adapted Multivariate SPC and Analytical Methods	Variable Selection Method	Variable Selection Approach Class	Process	Application	Step in SPC monitoring	Method Structure
2000-2005	Tur et al. (2002)	VS in FIR Qualitative Modeling	FIR	Multiple correlation coefficients PCA (B2 method) Cluster Analysis	Filter (Preprocessing)	Batch	Steam Generator	Detection	Framework
	Chu, Lee and Han (2004)	PLS *via* BBGVS	PLS Regression	SFFS (search algorithm) and minimization of the RMSEP from a multiple linear regression (selection criterion)	Filter (Preprocessing)	Batch	Chemical Industry	Detection	Framework
	Gourvénec, Capron, and Massart (2004)	GA applied to OPA	OPA	GA	Wrapper	Batch	Chemical Industry	Detection	Framework
	Zarzo and Ferrer (2004)	U-PLS integrated with VS and Block-wise PCR integrated with VS	U-PLS PCR	Technical knowledge of the process	Filter (Postprocessing)	Batch	Chemical Industry	Diagnosis	Framework
	Chiang, Pell, and Seasholtz (2004)	GA incorporated with FDA	FDA	GA	Wrapper	Continuous	Chemical Industry	Isolation	Framework
	Chu, Qin and Han (2004)	SVM integrated to entropy-based VS	SVM	SFFS (search algorithm) and entropy concept (selection criterion)	Filter (Preprocessing)	Batch	Semiconductor Industry	Detection	Framework
2006-2011	Wang and Jiang (2009)	VS-MSPC control chart	Generalized likelihood ratio test	FS	Filter (Preprocessing)	Continuous	Timber Industry	Isolation	Control Chart
	Yao et al. (2009)	2-D-DPCA with autodetermined Support Region	2-D-DPCA	Stepwise procedure AIC	Filter (Preprocessing)	Batch	Simulated Process	Isolation	Framework
	Zou and Qiu (2009)	LEWMA control chart	MEWMA	LASSO Regression LAR Regression	Filter (Preprocessing)	Continuous	Chemical Industry	Isolation	Control Chart

		Proposed Method (*continue*)	Adapted Multivariate SPC and Analytical Methods	Variable Selection Method	Variable Selection Approach Class	Process	Application	Step in SPC monitoring	Method Structure
2006-2011	Wang and Tsung (2009)	ADR-2 control chart	T^2 chart	MSN index	Filter (Preprocessing)	Continuous	Simulated process	Isolation	Control Chart
	González and Sánchez (2010)	Two-stage procedure for selection and evaluation of variables	PCA	Oblique rotation method	Wrapper	Continuous	Automotive Industry	Detection	Framework
	Ge, Zhang, and Song (2010)	BSPCA method	PCA	Subspace contribution index	Filter (Preprocessing)	Continuous	Chemical industry	Isolation	Control Chart
	Capizzi and Masarotto (2011)	LAR-EWMA control chart	EWMA	LAR	Filter (Preprocessing)	Continuous	Semiconductor manufacturing	Detection	Control Chart
	Ge, Gao, and Song (2011)	TDB method	PCA	Weight index Correlation analysis	Filter (Preprocessing)	Continuous	Chemical industry	Detection	Control Chart
2012-2017	Jiang, Wang, and Tsung (2012)	VS-MEWMA control chart	VS-MSPC control chart MEWMA control chart	FS	Filter (Preprocessing)	Continuous	Footwear Industry	Isolation	Control Chart
	Jeong et al. (2012)	Heuristic recursive VS method based on PCA and FA	PCA	Heuristic recursive VS method using FA	Wrapper	Continuous	Energy Industry	Detection	Framework
	Zou, Ning, and Tsung (2012)	LEWMA control chart for multivariate linear profile monitoring	MEWMA	LASSO Regression LAR Regression	Filter (Preprocessing)	Continuous	Logistics service	Isolation	Control Chart
	Ghosh, Ramteke, and Srinivasan (2014)	NSGA-II-JG based VS scheme	PCA	GA	Wrapper	Continuous	Chemical Industry	Detection	Framework
	Giannetti et al. (2014)	Co-linearity index to analyze mixed data	PCA	Co-linearity Index Graph Individual Penalty Matrix Approach Interaction Individual Penalty Matrix Approach	Filter (Postprocessing)	Continuous	Metallurgical Industry	Diagnosis	Framework
	Yan and Yao (2015)	Reconstruction-based fault isolation method using LASSO	Reconstruction-based approach	LASSO Regression LAR Regression	Wrapper	Continuous	Chemical Industry	Isolation	Framework

		Proposed Method (*continue*)	Adapted Multivariate SPC and Analytical Methods	Variable Selection Method	Variable Selection Approach Class	Process	Application	Step in SPC monitoring	Method Structure
2012-2017	Nishimura, Matsuura, and Suzuki (2015)	AIC-MEWMA control chart	VS-MEWMA control chart	FS AIC	Filter (Preprocessing)	Continuous	Metallurgical Industry	Isolation	Control Chart
	Kuang, Yan, and Yao (2015)	LASSO-based method EN-based method	FDA	LASSO Regression Ridge Regression LAR Regression	Wrapper	Continuous	Chemical Industry	Isolation	Framework
	Jiang, Yan, and Huang (2016)	FBPCA process monitoring method	PCA Contribution plots	GA	Wrapper	Continuous	Chemical Industry	Isolation	Framework
	Zhao and Wang (2016)	Faulty VS applied to reconstruction modeling	Reconstruction-based approach	Recursive VS method based on PCA decomposed subspaces	Wrapper	Continuous	Chemical Industry	Isolation	Framework
	Jiang and Huang (2016)	Distributed process monitoring framework	PCA	GA	Wrapper	Continuous	Chemical Industry	Isolation	Framework
	Li et al. (2017)	SR-VSMEWMA control chart	VS-MEWMA control chart SREWMA control chart	FS	Filter (Preprocessing)	Continuous	Food Industry	Isolation	Control Chart
	Abdella et al. (2017)	VS-MCUSUM control chart	MCUSUM	Stepwise procedure	Filter (Preprocessing)	Continuous	Hexagonal bolt manufacturing	Detection	Control Chart
	Shinozaki and Iida (2017)	VS based T^2 test	T^2 test	Estimate power *p*-value	Filter (Preprocessing)	Continuous	Simulated process	Detection	Framework
	Yan, Kuang, and Yao (2017)	SPLS-based fault isolation method	PLS regression	LAR algorithm	Wrapper	Batch	Injection moulding process	Isolation	Framework
	Zhao and Gao (2017)	SDISSIM algorithm for online incipient fault diagnosis	DISSIM method	Sparse regression LASSO Regression LARSEN algorithm	Wrapper	Continuous	Cigarette manufacturing	Isolation	Framework

2.4.1 Preprocessing filter approach

Research in this class is divided among authors who (*i*) integrated VS in Fuzzy Inductive Reasoning (FIR) methodology, (*ii*) applied PLS modeling preceded by a Bootstrapping-based Generalized Variable Selection (BBGVS) approach, (*iii*) used Support Vector Machine (SVM) pattern classification method integrated with entropy-based VS, (*iv*) improved the two-dimensional dynamic principal component analysis (2-D-DPCA), (*v*) applied VS based power estimate, and (*vi*) applied VS methods previous to the construction of MCCs.

Aiming at improving the performance of the FIR methodology, Tur et al. (2002) evaluated the integration of several VS algorithms to it. FIR qualitative modeling is used for predicting the trajectory behavior of measured variables, for control purposes. Techniques that target the elimination of variables with strong cross-correlation to other inputs (e.g. multiple correlation coefficients, PCA (B2 Method), and cluster analysis) performed considerably better than the method of the unreconstructed variance for the best reconstruction, and methods based on regression coefficients (ordinary least squares, PCR, and PLS), since they are more aggressive in discarding variables and provide faster convergence.

Chu, Lee and Han (2004) applied BBGVS as a Preprocessing step in PLS regression. Industrial data information were organized in an unfolded two-way matrix with process variables that could be related with the performance of quality variables inspected in the final product. Thirty different sets of bootstrapped data were obtained from the two-way matrix. A Sequential Forward Floating Selection (SFFS) was carried out to select variables to be included in each set; minimization of the Root-Mean-Square Error in Prediction (RMSEP) from a multiple linear regression was used as selection criterion. The frequency of selection of unfolded variables was selected in the various sets of bootstrapped data indicated those to be used as predictors in the PLS quality estimation model.

In another work, Chu, Qin and Han (2004) proposed the integration of SVM to entropy-based VS. The method was implemented in two phases. In the first phase, VS was performed using an entropy measure and the SFFS algorithm was used to determine variables that minimized the total entropy (assuming that larger entropy values indicate a higher degree of disorder in a dataset). The set of variables that minimize total entropy compose a hyperspace in which different data clusters are identifiable. After selecting variables, SVM classifiers were constructed to define decision boundaries between data clusters. Using the pattern classification method on the clustered dataset, correct boundaries between normal and fault data groups, and between different normal modes may be obtained, without relying on the normality assumption. Such information is used as criteria in the second phase, in which a hierarchical fault detection and operation mode identification takes place. To use the proposed method, data class information must be known *a priori*.

The 2-D-DPCA modeling method combines lagged regression and PCA to capture both the 2-D dynamics and cross-correlation information among process variables and lagged variables in batch processes. A key step in the method is the proper choice of a region of support (ROS) in which all lagged measurements should be located. Yao et al. (2009) proposed a method for ROS auto determination. First, a past neighborhood of the current sample is chosen as the candidate region of ROS, using prior process knowledge or through simple regression. A stepwise elimination is then iteratively carried out. In each run, a regression model is built to relate the remaining candidate independent variables to the current sample's value; models are evaluated using the Akaike information criterion (AIC) index. Then, one independent variable is eliminated from the candidate region based on the importance of variables calculated in each run. The best choice of the ROS is determined comparing index values calculated at each iteration. Once every variable's support region is determined, the combination of them will be the proper ROS to be used in the 2-D-DPCA model building. The SPE statistic and corresponding control limits may be calculated based on model residuals and used for online monitoring.

The last framework classified as Filter Preprocessing was proposed by Shinozaki and Iida (2017), that formulate the problem of detecting abnormal items as a hypothesis test based on the T^2 statistic. A VS method is used to maximize the test's power, i.e. the probability of detecting an abnormal item. From a reference sample of observations from the abnormal population (composed of abnormal items), subsets of variables are chosen and the power of the T^2 tests based on the subsets are estimated. The subset with maximum estimated power is the one containing the variables to be selected. Multiple subsets may have the same estimated power, especially when the number of abnormal items is not large. In those cases, the test's p-value is proposed as second criterion to determine the best subset, such that small values are preferred.

The integration of VS methods and MCC was first proposed by Wang and Jiang (2009), which developed the variable selection-multivariate statistical process control (VS-MSPC) control charts (CC). A new monitoring statistic derived from the generalized likelihood ratio test for a hypotheses test was proposed. By application of penalties and constraints to the equation that describes the rejection region for the null hypothesis, it was transformed into a penalized least squares problem in which μ_t was the coefficient vector to be estimated. To reduce the extensive computations needed to reach the optimal solution vector μ_t^*, a FS algorithm was implemented to select variables, such that the number of retained variables should be less or equal to parameter s, which is defined based on a priori knowledge of process experts and gives the maximum number of selected variables. The VS-MSPC chart statistic was obtained applying the optimal solution vector μ_t^* in the equation for the rejection region of the null hypothesis. When implementing this chart, the VS step identified potentially OOC variables and estimated their corresponding shift magnitudes; only the selected variables were included in the VS-MSPC chart. Whenever the chart triggered an OOC alarm all variables identified as potentially OOC were considered responsible for it, concluding the fault isolation.

Pursuing improvements in the performance of the VS-MSPC chart, Jiang, Wang, and Tsung (2012) proposed inserting a smoothing parameter in the penalized least squares equation, which led to the proposition of the VS-multivariate exponentially weighted moving average (MEWMA) CC. Similarly to the VS-MSPC chart, the solution was obtained using a FS algorithm and a stopping parameter s. To obtain the VS-MEWMA chart statistic, the optimum solution vector μ_t^* was applied in the VS-MSPC chart statistic and the EWMA statistic w_t replaced the p-dimensional measurement vector y_t, observed at time t, in the monitoring equation aiming at improving the method’s sensitivity. The chart triggers an alarm when the VS-MEWMA chart statistic is higher than an upper control limit chosen for a desired performance.

Seeking improvements in the VS-MEWMA chart two other methods were developed. First, Nishimura, Matsuura, and Suzuki (2015) proposed the Akaike information criterion-multivariate exponentially weighted moving average (AIC-MEWMA) CC that uses a new criterion to determine the value of parameter s aiming to improve the constrained optimization step. In their proposition the AIC is used to define the minimum number of variables to be retained in the VS step. Recently, Li et al.

(2017) proposed the self-starting spatial rank multivariate EWMA CC using forward variable selection (SR-VSMEWMA), which integrates the multivariate spatial rank and FS in an EWMA scheme to monitor processes with sparse mean shifts. Primarily, a self-starting technique was applied to the VS-MEWMA chart, and μ_0 and Σ at the current time point t were replaced by appropriate estimators constructed from previous observations, allowing reduction of the required IC samples. The new charting statistic is not transformation invariant, so spatial rank was carried out to transform the original data and guarantee that the distribution of the resulting charting statistic was fixed, regardless of IC parameters. The transformed data were combined with the VS-MEWMA modified by the self-starting technique, originating the robust SR-VSMEWMA.

Incorporating the LASSO VS method into the SPC problem, a new CC was proposed by Zou and Qiu (2009) for monitoring multiple parameters. That CC was later improved by Zou, Ning, and Tsung (2012) to monitor general multivariate linear profiles. Zou and Qiu (2009) used the sparsity property of the LASSO method to select the exact set of nonzero regression coefficients in multivariate regression modeling and propose a LASSO-based multivariate test statistic. The statistic was integrated in a MEWMA charting scheme for online multivariate process monitoring. The result was the proposition of a LEWMA CC based on the Adaptive LASSO penalized likelihood. The new CC was able to detect possible shift directions automatically, each time a new vector of observations was made available. Once the CC triggers a mean shift, the shift location is estimated and the specific measurement components that caused the shift are identified. Shift location is estimated through the generalized maximum likelihood approach for change-point detection; shift components were identified through the LASSO methodology choosing one of the LASSO estimators using a model selection criterion (e.g. risk inflation criterion). Since some estimators' components are exactly zero, those that differ from zero are deemed responsible for the shift with no need for any extra tests, which are commonly required in most existing fault isolation methods. Zou, Ning, and Tsung (2012) extended the LEWMA chart using in a single CC both coefficients and variances of a multivariate linear profile.

Aiming to develop a CC with a broader scope that could handle "unstructured" cases, profiles and multistage processes, Capizzi and Masarotto (2011) developed the LAR-EWMA CC. The MCC is able to detect shifts in the mean and increases in process dispersion. As in Wang and Jiang (2009), and Zou and Qiu (2009), the authors propose the monitoring of subsets of possible OOC variables. To achieve that, the Least Angle Regression (LAR) algorithm was integrated with the MEWMA CC. Briefly, LAR starts with all coefficients set to zero and then proceeds in h successive steps; in each step a potential predictor is added to the model. Once a predictor is selected by LAR in the i-th step, its set of coefficients is viewed as a promising k-dimensional set of parameters for which a shift may have occurred. So, for $k = 1, \ldots, h$, an alternative hypotheses should be formulated and the corresponding generalized log-likelihood ratio (GLR) statistic computed. GLR estimates coefficients and LAR constrains to zero those of variables not selected by LAR during the first k steps of the procedure. To also detect increases in dispersion, an additional alternative hypothesis should be formulated, along with its related one-side EWMA statistic. Eventually, the overall statistic is computed for monitoring.

So far, authors have worked with EWMA-based CCs; some other CCs are proposed in the following articles. Abdella et al. (2017) expanded the concept and integrated VS procedures to a multivariate cumulative sum (MCUSUM) CC, proposing the variable selection-based multivariate cumulative sum (VS-MCUSUM) CC to improve performance in the detection of small mean changes in process parameters. Similar to the VS-MSPC (WANG; JIANG, 2009) and VS-MEWMA (JIANG; WANG; TSUNG, 2012) CC, the proposed method uses a VS algorithm to identify a subset of process variables possibly affected by the presence of assignable causes; only such variables are continuously monitored. In the VS-MCUSUM CC, a stepwise VS is adopted and the F-ratio test used to identify the set of variables most likely to cause process changes. The procedure stops when q variables are selected, such that q is a parameter set by an experienced quality practitioner that represents the number of changed variables. The dimension of the mean vector $\mathbf{y}_t$ is reduced to q, and used to calculate the value of the cumulative sum (CUSUM) statistic.

Focusing on the monitoring of nonlinear processes, Ge, Zhang, and Song (2010) developed the Bayesian subspace-PCA (BSPCA) method. In their proposition, the original nonlinear space is initially approximated by several linear subspaces through PCA decomposition in the principal component and in the residual subspaces. Then, in each linear subspace the subset of most relevant variables is selected using two new subspace contribution indices. Next, subspace monitoring models are constructed based on the selected subsets, and confidence limits of their corresponding monitoring statistics are determined. For each monitored sample, monitoring results from different linear subspaces are combined using Bayesian inference, which transforms traditional monitoring statistic values into fault probabilities in each individual subspace, to allow the combination of results from different subspaces. With that, new monitoring T^2 and SPE CCs are generated to detect process abnormalities. Once a fault is detected, a newly proposed fault isolation approach is implemented using the reconstruction-based contribution plot method on each linear subspace. Subspace results are finally combined to yield a decision. Both fault isolation and magnitude may be obtained simultaneously.

Extending the procedure above, Ge, Gao, and Song (2011) developed the two-dimensional Bayesian (TDB) monitoring method for nonlinear multimode processes. Briefly, a process dataset is partitioned using k-means clustering, rendering a multiple sub-group dataset corresponding to different operation modes. Each sub-group dataset is further partitioned into several linear subspaces. Different from Ge, Zhang, and Song (2010), VS is conducted using a two-step strategy. First a weighted index is implemented to select a subset of most important variables in the linear subspace. Then, correlations between each variable and remaining variables in the selected subset are evaluated, such that variables with large sums of correlation values are selected for linear subspace construction. A PCA model is developed in each linear subspace, for different operation modes. For online monitoring of new data samples, T^2 and SPE chart statistics are calculated in each linear subspace. To combine monitoring results from different operation modes, a two-dimensional Bayesian monitoring approach is applied. First, posterior probabilities of each operation mode are determined and then Bayesian inference is employed to determine fault probabilities. Finally a fault detection index is calculated for each linear subspace and combined in a final nonlinear fault detection index (NFDI). Whenever NFDI values are above the confidence limits, some fault is considered to be acting on the process.

Closing the application of Preprocessing filter VS approaches to MSPC, Wang and Tsung (2009) proposed a new CC to monitor processes with dynamic mean shifts. Two adaptive dimension reduction (ADR) charts are proposed, being the ADR-2 chart suitable to high dimensional datasets. In the ADR chart, the projection matrix performs in such a way that the contribution of each variable is evaluated at each step, and redundant variables are abandoned dynamically. Independent components are obtained via orthogonal decomposition using Mason, Tracy and Young (MYT)'s decomposition of the T^2 value. In order to fit the MYT-decomposed components into the projection framework, three projection matrices are defined. Alarms are interpreted according to the projection matrix issuing them, as follows: (*i*) first matrix: signal originated in the input stream; (*ii*) second matrix: signal originated in the conditional output, when input status is known to be IC; (*iii*) third matrix: signal leads to the monitoring of the original vector, suggesting process failure. To choose the best combination of MYT decomposed components, VS is conducted based on the multivariate standardized shift (MSN) index, which measures the performance of the T^2 chart: components are therefore added or dropped based on their contribution to the MSN.

2.4.2 Postprocessing filter approach

Research in this class is divided among authors who applied VS based (*i*) on expert knowledge about the process, and (*ii*) on the co-linearity index.

Zarzo and Ferrer (2004) proposed filtering process variables that are most correlated with a final quality parameter. For that, two methods were proposed. The first used Unfold Partial Least Square Regression (U-PLS) with progressive simplification of the model through technical knowledge to define the causal correlations. Trajectories of PLS weights were analyzed by juxtaposition of unfolded variables, in order to distinguish groups (denoted as "correlation runs") that were related to process deviations. Once correlation runs were identified they were matched with the original trajectory, and technical knowledge was applied in search of a diagnosis, or to find an explanation for the observed correlation. Whenever no reasonable explanation was found and the process variable in the period with correlation was considered non-important, the entire trajectory was removed from the dataset. The method was considered adequate for the purpose of diagnosis, since it retained the main variables that contributed to the model prediction capacity. The second method, named Block-wise Principal Component Regression, was proposed considering each variable trajectory as a block. Carrying out PCA in the initial unfolded matrix, a Principal Component score matrix was obtained and analyzed using simple linear regression in search of predictive models for the process final quality variables. Each predictive model was evaluated with respect to two parameters to reduce the number of variables: the squared linear correlation coefficient and the *p*-value. Next, expert knowledge was used to promote a VS by comparing the CUSUM CC of each resulting variable with the CUSUM CC of the response variable. Even though the causes of variability of the response variable have not yet been identified, some process variables were pointed out as likely to be critical. To finish the diagnosis, designed experiments should be run with those variables to define optimal process settings that would minimize process variability and improve its final quality.

Searching for the best monitoring parameters for a specific industrial process, Giannetti et al. (2014) adapted the method that combines the co-linearity index and the penalty matrix approach, originally proposed by Ransing et al. (2013). To promote VS using this method, a two-dimensional co-linearity index plot was constructed for each pair of response and process variables by drawing a vector with the dimensions of the PCA loadings that provided a reduced representation of those variables. Next, noise-free correlations between IC process variables and response variables were quantified and visualized, such that the higher the magnitude of a process variable, the largest its importance in describing the dataset variance. VS was carried out directly in the plots. Once selected, variables were analyzed using penalty matrices for each variable and their interactions; such matrices converted hypotheses raised in the VS step in process information. Extend this strategy to analyze mixed data composed of continuous and categorical variables, Giannetti et al. (2014) proposed a robust method for pre-treating data based on Multiple Factor Analysis. The original dataset was reorganized in three main groups of variables (response, categorical, and quantitative variables). Separate analyses were carried out in each group, variables were redistributed and the different groups were merged back into a single dataset, which was then analyzed using the co-linearity index method described above.

2.4.3 Wrapper approach

Works in this class proposed the use of oblique rotation, GA, factor analysis (FA), and LASSO regression and elastic net (EN) regularization to promote VS.

To select and evaluate variables that should be used in a MSPC, González and Sánchez (2010) proposed a two-stage iterative procedure. In the first stage, the oblique rotation method was applied to select the single variable that carried the largest amount of information in the original set of variables. To start the method, a Varimax rotation was applied to factors obtained through PCA; the rotation was then extended to an oblique solution *via* Promax. The rotated component with the maximum sum of squared loadings was identified, and the selected variable was the one with the largest absolute loading in that component. In the second stage, the selected variable was evaluated following two approaches. The first approach was based on R-like indices that informed the amount of residual information in the variables not selected; the second approach was based on the T^2 chart average run length (ARL) when only the selected variable was used to evaluate the performance in the detection of simulated OOC events, comparing the result with the ARL when using all variables. If results in the evaluation step were considered satisfactory by the analyst, the iterative VS method was stopped; otherwise, the method was restarted removing the information carried by selected variables from those not yet selected.

The works of Gourvénec, Capron, and Massart (2004), Ghosh, Ramteke, and Srinivasan (2014), Jiang, Yan, and Huang (2016), and Jiang and Huang (2016) used GA coupled with PCA. The work of Chiang, Pell, and Seasholtz (2004) coupled GA with FDA.

Gourvénec, Capron, and Massart (2004) proposed an improvement in the orthogonal projection approach (OPA) described by Gourvénec et al. (2003) for monitoring batch processes. When OPA is applied to data from a chemical mixture process it is possible to define the ideal number of components in the mixture and find the set of pure spectra deemed representative of the process. That allows the online estimation of concentration values of new batches every time a new spectra is recorded. Gourvénec, Capron, and Massart (2004) proposed coupling GA and OPA to obtain smaller spectra through the selection of ranges of Near infrared (NIR) wavelengths, reducing the time to acquire and transfer spectra information to the database. Briefly, the initial population was generated randomly and represented the number of possible candidate solutions. The evaluation of solutions was carried out based on the measure of dissimilarity between concentration profiles obtained with all variables, and with a subset of selected variables. Once the initial population had been evaluated, it evolved to yield new solutions using the genetic operators of reproduction and mutation, and the optimal solution was obtained when the dissimilarity was minimized. Crossover and mutation was performed as two independent steps, measuring the dissimilarity of the profiles after each step. At each iterative step, the GA selected the wavelengths and generated a reduced spectra able to speed up the process and preserve the critical information in the original database, yielding better results for OPA.

The second method that uses GA to select variables, due to Ghosh, Ramteke, and Srinivasan (2014), proposed a reduced PCA model to optimize the monitoring performance in a multi-fault setting. A VS scheme based on the non-dominated sorting genetic algorithm and a jumping gene operator (NSGA-II-JG) was proposed to reduce the number of variables to be monitored through the identification of a subset of variables that minimizes the cumulative error given by the sum of two error rates: False Alarm Rate (FAR) and Missed Detection Rate (MDR). Each chromosome represented a subset of variables selected from the normal multivariate training dataset and the confidence limits for T^2 and SPE statistics to be applied to the validation data, to evaluate the performance of the PCA monitoring model. After validation, chromosomes were submitted to four genetic operators (selection, crossover, mutation, and jumping gene), and the population of the next generation was obtained by elitism. GA terminated when the maximum number of generations was reached.

The third method using GA for VS was due to Jiang, Yan, and Huang (2016). It proposed the Fault-bayesian PCA (FBPCA) process monitoring method aiming at the improvement of an NSGA-II-JG-based VS scheme (GHOSH; RAMTEKE; SRINIVASAN, 2014). The method used GA to select the optimal variables and developed a specific reduced PCA model for each fault. Assuming that there were b faults in the validation set and that the remaining variables were assigned to a single block, $b + 1$ blocks existed at the end of the procedure, and for each one a reduced PCA was constructed. Then, similar to the former method, monitoring results for each sub-block were obtained by computing FAR and MDR, and GA continues until the best possible performance is achieved for one specific fault, or the stop rule is reached. After that, a Bayesian inference fusion scheme evaluated all subsets and constructed the final monitoring statistics. If a fault was detected, an FBPCA contribution plot was used to isolate the variables and determine their contribution in a specific block. The total contribution of a variable, considering that it can contribute in more than one block, should be calculated as the sum of its weighted contributions.

The last proposition using GA coupled to PCA is the distributed process monitoring framework by Jiang and Huang (2016). They start by dividing all measured variables into M blocks. To achieve the best possible monitoring performance from a process decomposition perspective, a GA-based performance-driven process decomposition method, with a user-determined number of sub-blocks, is implemented. The objective function for the GA-based optimization aims at minimizing the MDR. In GA-optimization an initial population of chromosomes is randomly generated, and variables are then divided into sub-blocks. Based on temporary block division results, local PCA monitoring may be established and the value of the fitness function calculated. Once the final chromosome is obtained, variables are divided into sub-blocks and a PCA monitoring model is established for each sub-block. T^2 and SPE statistics are constructed for each new sample from a m^{th} sub-block. These statistics are combined in a Bayesian Inference Comprehensive statistic that is used for fault detection. To promote fault isolation, the authors adapted the Bayesian fault isolation method for centralized monitoring by Jiang, Huang, and Yan (2016) to handle distributed monitoring, redefining the objective function to minimize the MDR.

Chiang, Pell, and Seasholtz (2004) incorporated a GA to FDA for process fault identification. As will be further discussed when presenting Kuang, Yan, and Yao (2015)'s proposition, the basic assumption is that variables can be discriminated in two classes, normal and faulty. The method starts by randomly creating chromosomes composed by different subsets of the original variables. The performance of each chromosome is evaluated using a leave-1/5-out cross validation scheme with FDA, and the fitness function is calculated for all chromosomes. Cross-over and mutations are performed over the evolutions to increase the fitness function, improving chromosomes. At the end of evolutions, the chromosome with highest fitness function is saved. The procedure is repeated iteratively and the final chromosome with the highest fitness function, after all evolutions, is saved. At the end, several chromosomes are retained. A bar chart of the frequency of selection of each variable is then constructed. Variables are sorted according to their frequency of selection, and the number of variables required to explain the shift is determined by maximizing the fitness function.

Jeong et al. (2012) proposed a heuristic recursive VS method based on FA to improve PCA modeling. First PCA was applied to normal operation data. When the cumulative sum of the explained variance retaining the first two principal components was higher than 80%, T^2 and SPE statistics were run for validation. Otherwise, FA rearranged the variables in a descending order of standardized score coefficients, to group those located in a similar process region. The iterative process was carried out until the criteria was satisfied, and a comparative evaluation was made using the PCA score plot of each group of variables.

The reconstruction modeling for fault isolation proposed by Zhao, Sun and Gao (2012) was improved by Zhao and Wang (2016) through VS of the most significant OOC variables for each fault. First, PCA-based monitoring models were constructed. When an alarm was triggered, the effects of faults were decomposed in the principal component subspace (PCS) and residual subspace (RS) revealing the most important directions of fault deviations relative to normal process conditions, which were used to reconstruct the principal fault systematic deviations. Significance of variables was evaluated by a quantitative statistical index named reconstruction-based variable contribution. The corrected part of the fault was then checked, process variables were sorted by the mean variable contributions along the time direction for multiple samples, and the variable with the largest mean value was determined to be the most informative and contributive. That variable was stored in the faulty variable library, and removed from both the normal and OOC datasets. The updated normal dataset was used to redevelop the PCA monitoring models. If the faulty variables were assumed to be normal, the procedure should be stopped; otherwise, if there were 10 consecutive monitoring alarms the procedure was recursively repeated until all alarm-relevant variables were selected. The final result was a subset of OOC variables that were relevant regarding the alarms of monitoring statistics. For each type of fault there were a subset of significant OOC variables; variables not selected in any subset were deemed irrelevant regarding fault isolation. Finally, a parsimonious reconstruction model for fault isolation was built based on the selected OOC variables.

A drawback in Zhao and Wang (2016)'s method is the assumption that a sufficient historical fault database, will be available, which may not always be the case in practice. In addition, it may be difficult to handle unknown disturbances not covered by historical fault data. To overcome that, Zhao and Gao (2017) proposed the sparse dissimilarity (SDISSIM) algorithm to identify the incipient variables that are responsible for the changes of distribution structure without a priori fault information. The DISSIM method (KANO; HASEBE; HASHIMOTO, 2002) considers that distribution variances may be used to represent the distribution dissimilarity between two datasets. For that, it quantitatively evaluates the distribution difference between normal and faulty conditions calculating the difference between process variances. SDISSIM extends this concept to fault isolation of abnormal variables that distort the variable covariance structure. First the dissimilarity distribution is decomposed and the critical dissimilarity component is extracted. Next, a sparse regression-type optimization is run to obtain sparse coefficients using LASSO regression and isolate the fraction of variables deemed abnormal. Whenever the sample dimension is smaller than the variables' dimension, EN will be used to construct the optimization problem. Whenever a variable is removed the remaining variables are compared with respect to normal and faulty cases by rebuilding a new reference model for the remaining variables under normal conditions, and checking whether remaining variables operating under faulty conditions behave similarly to those in the normal case. If they are deemed similar, it means that all faulty variables have been removed; if not, another variable should be removed. Carrying on this iterative procedure it is possible determine the proper number of faulty variables to be retained.

Closing the methods which integrate MSPC with wrapper VS approach, three propositions used LASSO regression to improve fault isolation (KUANG; YAN; YAO, 2015; YAN; KUANG; YAO, 2017; YAN; YAO, 2015). In the first, Yan and Yao (2015) improved the reconstruction-based approach presented by Yue and Qin (2001) through the proposition of a new graphical method for fault isolation based on PCA. The insertion of a regularization parameter in the reconstruction equation was proposed to approximate mathematically the method to the LASSO regression algorithm, and allow the identification of fault directions and the selection of variables responsible for each type of fault. Once a fault occurred, the subspace that characterized it was identified, and for each type of fault a regularization parameter was assigned to represent the fault's change direction and the transition point between faults. An adapted LAR algorithm was used to define regularization parameter values and to promote the sequential estimation of coefficients, adding one by one to the model to compose the active set. Once convergence was achieved, variables in the active set were considered as potentially responsible for a specific fault; these variables were then reconstructed, tested in the Combined Index (CI) monitoring statistic, that uses T^2 and SPE statistics simultaneously, and compared to the statistic applied to the original data. As the CI monitoring statistic returns to the IC situation after several variables were reconstructed, such variables were identified as the ones most related to the fault.

Methods proposed by Kuang, Yan, and Yao (2015) were based on the assumption that the fault isolation task could be considered as an instance of a discriminant analysis in which the variables were assigned to a normal operating data class or to a class of data associated with the detected fault. When the right choice of predictors and response variable is made, FDA becomes identical to the least squares regression model, and the problem of multivariate fault isolation could be formulated as a penalized regression. By the introduction of an L_1 regularization in the standard multiple regression model, VS could be achieved through a LASSO-based model. Instead of identifying OOC variables based on control limits computed using normal operating data, the proposed method provides a sequence of process variables according to their relevance to the detected faults: the earlier the variable appears in the active set, the more significantly it relates to the fault. The method above does not handle properly highly correlated faulty variables, and may not identify all OOC variables. The second method proposed by Kuang, Yan, and Yao (2015) handles that drawback using an EN regularization technique. The LASSO-based method was revised by adding a L_2 penalty term in the objective function of least squares regression. As in the first proposed method, results are presented as a sequence of process variables entering the active set. The EN-based method is less likely to misidentify the strongly correlated faulty variables.

Yan, Kuang, and Yao (2017) proposed the Sparse PLS (SPLS)-based fault isolation method to handle autocorrelations and cross-correlations that characterize batch process data. SPLS is based on the equivalence between fault isolation and VS in a two-class discriminant problem, demonstrated by Kuang, Yan, and Yao (2015). SPLS builds a discriminant analysis model for normal and faulty operation batch data, achieving modeling and VS simultaneously. The model is adjusted gradually, such that variables enter the active set sequentially, reflecting their importance in characterizing process abnormalities and based on the concept of transition points, described in Yan and Yao (2015). The order in which variables enter the active set reflects not only their importance, but also indicates the most critical time interval for detecting batch abnormalities.

2.5 PROCESS MONITORING IN SPC AND PERFORMANCE OF THE DEVELOPED METHODS

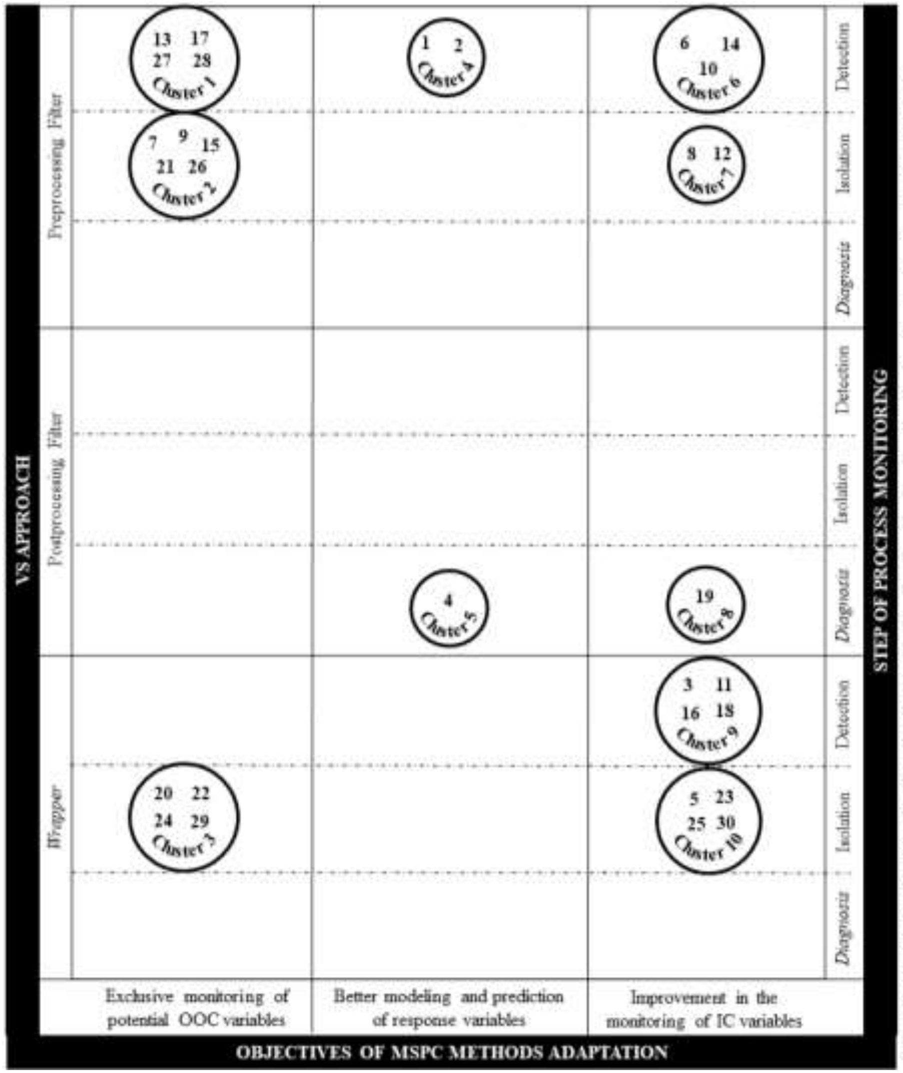
Process monitoring in SPC is carried out in two phases. In Phase I data are collected, the process stability is evaluated, and an appropriate monitoring IC model is developed. In Phase II, such monitoring model is implemented with data collected successively over time to identify abnormal process behaviors (fault detection). Then process variables contributing most to the detected fault are identified (fault isolation), and the root cause of the observed OOC status are determined (fault diagnosis). Finally, fault effects are removed from the data (process recovery). The capability to detect process faults quickly, which addresses the sensitivity of an MSPC scheme, and the ability to locate shifted variables accurately, that concerns the diagnostic capability of the scheme, are great challenges in MSPC process monitoring (JIANG; WANG;

TSUNG, 2012; KUANG; YAN; YAO, 2015; WOODALL; MONTGOMERY, 2014). With that in mind, methods in this review were analyzed to address our third research question (*'which steps of process monitoring in SPC were studied?'*). Fault detection and fault isolation in Phase II were studied by 13 and 15 methods, respectively; fault diagnosis was the subject of two methods.

In the current section we categorize the methods classified in section 2.4 according to (*i*) their objectives (presented in section 2.3), and (*ii*) the step of process monitoring they address; that led to the ten clusters of methods presented in Figure 2.3. Half of the clusters comprised 70% of the studied methods; they are: Filter Preprocessing VS approach to the exclusive monitoring of potential OOC variables to improve fault detection and isolation, Wrapper VS approach to improve the monitoring of IC variables as well as fault detection and isolation, and Wrapper VS approach to exclusively monitor potential OOC variables and improve fault isolation.

Some objectives are predominant given the VS approach they are based on. That is the case of methods aimed at monitoring potential OOC variables, whose main VS approach is Filter Preprocessing, and methods aimed at improving the monitoring of IC variables, whose main VS approach is Wrapper. Choosing the best VS approach to achieve each objective is related to the desired properties of each method. Objectives of MSPC methods that used the Filter Preprocessing approach were centered at computational efficiency by discarding irrelevant or redundant variables before the application of MSPC. MSPC methods that used a Wrapper VS approach were more often targeted at improving the accuracy of MSPC through monitoring a reduced number of IC variables.

Clusters were sorted according to the objectives of MSPC adaptations, and will be discussed on the basis of the monitoring step they were aimed at, and the performance of the proposed methods against traditional MSPC methods.



METHODS

1 – VS in FIR Qualitative Modeling (TUR et al., 2002)

2 - PLS via BBGVS (CHU ; LEE ; HAN, 2004)

3 – GA applied to OPA (GOURVÉNEC; CAPRON; MASSART, 2004)

4 - U-PLS integrated with VS and Block-wise PCR integrated with VS (ZARZO; FERRER, 2004)

5 – GA incorporated with FDA (CHIANG; PELL; SEASHOLTZ, 2004)

6 - SVM integrated to entropy-based VS (CHU; QIN; HAN, 2004)

7 - VS-MSPC control chart (WANG; JIANG, 2009)

8 - 2-D-DPCA with autodetermined Support Region (YAO et al., 2009)

9 – LEWMA control chart (ZOU; QIU, 2009)

10 - ADR-2 control chart (WANG; TSUNG, 2009)

11 - Two-stage procedure for selection and evaluation of variables (GONZÁLEZ; SÁNCHEZ, 2010)

12 – BSPCA method (GE; ZHANG; SONG, 2010)

13 – LAR-EWMA control chart (CAPIZZI; MASAROTTO, 2011)

14 – TDB method (GE; GAO; SONG, 2011)

15 - VS-MEWMA control chart (JIANG; WANG; TSUNG, 2012)

16 - Heuristic recursive VS method based on PCA and FA (JEONG et al., 2012)

17 - LEWMA control chart for multivariate linear profile monitoring (ZOU; NING; TSUNG, 2012)

18 - NSGA-II-JG based VS scheme (GHOSH; RAMTEKE; SRINIVASAN, 2014)

19 - Co-linearity index for analyzing mixed data (GIANNETTI et al., 2014)

20 - Reconstruction-based fault isolation method using LASSO (YAN; YAO, 2015)

21 - AIC-MEWMA control chart (NISHIMURA; MATSUURA; SUZUKI, 2015)

22 – LASSO-based method and EN-based method (KUANG; YAN; YAO, 2015)

23 – FBPCA process monitoring method (JIANG; YAN; HUANG, 2016)

24 - Faulty VS applied to reconstruction modeling (ZHAO; WANG, 2016)

25 – Distributed process monitoring framework (JIANG; HUANG, 2016)

26 – SR-VSMEWMA control chart (LI et al., 2017)

27 – VS-MCUSUM control chart (ABDELLA et al., 2017)

28 - VS based T^2 test (SHINOZAKI; IIDA, 2017)

29 – SPLS-based fault isolation method (YAN; KUANG; YAO, 2017)

30 - SDISSIM algorithm for online incipient fault diagnosis (ZHAO; GAO, 2017)

Figure 2. 3 VS-MSPC integration approaches clustered according to objectives and step of process monitoring they address

2.5.1 Exclusive monitoring of potential OOC variables

The first 3 clusters were related to this objective, encompassing 13 methods. They used 6 different VS approaches and were aimed at improving fault isolation (77%) or fault detection (23%). Five methods in those clusters shared a characteristic: they used the L_0-norm, L_1-norm and L_2-norm penalty types to remove variables with estimated coefficients of small magnitude.

All 13 datasets used to illustrate propositions in this section were obtained from continuous processes, except for Yan, Kuang, and Yao (2017)'s, which was obtained from a batch process. In Cluster 1, CCs were applied in real industrial setups [semiconductor manufacturing in Capizzi and Masarotto (2011), and hexagonal bolt manufacturing in Abdella et al. (2017)], except for Shinozaki and Iida (2017), whose proposition was applied to simulated data. MCCs presented in Cluster 2 were applied to data from timber (WANG; JIANG, 2009), chemical (ZOU; QIU, 2009), footwear (JIANG; WANG; TSUNG, 2012), metallurgical (NISHIMURA; MATSUURA; SUZUKI, 2015) and food (LI et al., 2017) industries, and to data from a logistics service (ZOU; NING; TSUNG, 2012) attesting the potential and versatility of these new MSPC methodologies. Methods in Clusters 3 were applied to the Tennessee Eastman Process (TEP) (KUANG; YAN; YAO, 2015; YAN; YAO, 2015; ZHAO; WANG, 2016), which is a well-known benchmark simulation that provides realistic chemical industrial process data for evaluating process control and monitoring methods. The method proposed by Yan, Kuang, and Yao (2017) was applied to an injection moulding batch process dataset.

2.5.1.1 Fault Isolation

Methods developed to achieve fault isolation are positioned in Clusters 2 and 3. Those in Cluster 2 used FS and adaptive LASSO penalized likelihood to select variables, previous to the construction of five new MCCs. The isolation was carried out using variables with nonzero coefficients as basis for further identification of root causes. Those CCs that use FS seek to improve fault isolation by ensuring a better performance than Hotelling's T^2 chart in the context of high dimensional multivariate datasets. The VS-MSPC chart (WANG; JIANG, 2009) was considered superior to Hotelling's T^2 chart in detecting moderate and large shifts; the VS-MEWMA chart (JIANG; WANG; TSUNG, 2012) was superior to the T^2, MEWMA, and VS-MSPC

CCs in detecting small shifts, illustrating the increase in sensitivity due to the special weighing of recent observations from the process implemented in the chart statistic. Moreover, the VS-MEWMA was able to efficiently detect sparse shifts and was robust to inaccurate specifications in the value of parameter s. The AIC-MEWMA chart (NISHIMURA; MATSUURA; SUZUKI, 2015) displayed superior performance when only one or two variables shifted, probably due to the severe AIC penalty imposed when a larger number of variables was involved. Finally the SR-VSMEWMA stood out, since it could efficiently detect sparse shifts, especially when the process distribution is heavy-tailed or skewed, did not need prior knowledge of IC distribution, which makes it appropiate to start-up situations, and is robust to non-normally distributed data. The LEWMA CC (ZOU; QIU, 2009), based on LASSO, displayed chart statistic values much larger than its control limit when compared with REWMA and MEWMA CCs, better signalizing the occurrence of a shift. The LEWMA CC is suitable for cases in which knowledge of shift patterns is little or nonexistent; on the other hand, FS-based CCs are suitable for situations when the number of potential OOC is known *a priori*. The LEWMA CC showed to be capable of monitoring multivariate linear profile data (ZOU; NING; TSUNG, 2012). In comparison with MEWMA adapted to multivariate profiles, LEWMA provided reasonable diagnostic ability to identify shifted parameters in numerical simulated results. One important point is that LEWMA is affected by the size of the reference dataset.

Works in Cluster 3 used the wrapper approach for the purpose of fault isolation (KUANG; YAN; YAO, 2015; YAN; KUANG; YAO, 2017; YAN; YAO, 2015; ZHAO; WANG, 2016). Results obtained with faulty VS applied to reconstruction modeling (ZHAO; WANG, 2016) were compared with those using a progressive PCA algorithm showing that, even when both methods correctly identified a similar number of faulty variables, the proposed algorithm displayed a lower of variables wrongly picked up as faulty variables, resulting on a significantly better performance. This method made easier to distinguish among different faults, even when faults were divided among the same OOC variables. A reconstruction model without VS was also used as comparative for the online fault isolation performance demonstrating the superiority of the proposed method in correct fault isolation, particularly for the T^2 statistics. Also promoting a better fault reconstruction, the graphical method of Yan and Yao (2015) demanded much shorter computational time for isolating an abnormal sample using LASSO regression when compared to the reconstruction using the branch and bound algorithm.

The traditional T^2 and SPE contribution plots were used for comparison of fault isolation performance using LASSO-based and EN-based methods (KUANG; YAN; YAO, 2015). Contribution plots presented some false alarms during evaluation, which was not evidenced in these new methods showing their superior performance. In comparison to LASSO-based method, EN-based methods proved to be better to isolate highly correlated fault variables. Finally, SPLS based on discriminant analysis (YAN; KUANG; YAO, 2017) enabled the identification of the most critical variable in a detected fault in a batch process, through the visualization of which variable is first involved in the SPLS model as long as the shrinkage task occurs using LAR algorithm. This method outperforms MPCA-based contribution plots and PLS-DA, which did not identify faulty variables clearly.

2.5.1.2 Fault Detection

Fault detection using only OOC variables was the subject of three methods in Cluster 1. Two of them proposed new CCs; one proposed a framework. All methods analyzed the improvement in the detection of changes in the mean vector assuming that process dispersion did not change. The VS-MCUSUM CC in Abdella et al. (2017) showed significant advantage in shift detection under a wide range of process settings when compared with the traditional MCUSUM and T^2 CCs. The LAR-EWMA CC, due to Capizzi and Masarotto (2011), was capable to detect shifts in several representative OOC scenarios (e.g. in elements of the mean vector, and in profile and multistage monitoring) that were not detected by other EWMA-based CCs that do not use a VS algorithm. The VS procedure proposed for the LAR-EWMA CC only detects changes in the mean vector; changes in process dispersion are not contemplated. The framework proposed by Shinozaki and Iida (2017) showed that whenever the sample size from the population of abnormal items increases, the performance of VS also increases, improving the probability of detection of an abnormal item when a T^2 test is run on the selected variables. The performance of the method is dependent on the availability of a large dataset of abnormal items, which may be viewed as a drawback in several applications.

2.5.2 Better modeling and prediction of response variables

The two clusters comprised of methods developed to achieve better modeling and prediction of response variables include 3 methods to improve batch process monitoring, which use 3 distinct VS approaches.

2.5.2.1 Fault Detection

The two methods in Cluster 4 aimed at improving fault detection using a preprocessing filter approach. Tur et al. (2002) proposed one of the only methods that accepts qualitative input data and not only predicts the output variable, but provides a confidence measure for the prediction. When integrated to VS, the FIR qualitative modeling of a steam generator displayed reduced computational complexity yielding high predictability and specificity. The second method, due to Chu, Lee and Han (2004), was an alternative to multiway partial least squares (MPLS) in the analysis of a three-way dataset from a polymerization batch process. PLS via BBGVS showed superior prediction accuracy than MPLS, which could be attributed to the Filter Preprocessing approach that selected variables and, consequently, increased the correlation between process and quality variables. The drawback of the method was that the computational cost was higher than that of MPLS.

2.5.2.2 Fault Diagnosis

The method by Zarzo and Ferrer (2004) in Cluster 5 used technical knowledge followed by a planned experiment to diagnose the critical points of a polymer production batch process. In addition to deep process knowledge, the method requires careful analysis of CCs and variables’ trajectories, which is time consuming.

2.5.3 Improvement in the monitoring of IC variables

Fourteen of the thirty papers covered in this review used VS with the objective of improving the monitoring of IC variables; they were assigned to five clusters (Clusters 6 to 10). Seven of them were developed for fault detection, 6 proposed improvements in fault isolation, and 1 was targeted at fault diagnosis. A total of 10 VS approaches were proposed.

Three of the proposed methods that adapted MSPC strategies were applied to batch processes, either using real data from industry [chemical in Gourvénec, Capron, and Massart (2004), and semiconductor in Chu, Qin and Han (2004)] or from a simulated process (YAO et al., 2009). The remaining eleven methods were developed for application in continuous processes. The methods proposed by Chiang, Pell, and Seasholtz (2004), Ge, Zhang, and Song (2010), Ge, Gao, and Song (2011), Ghosh, Ramteke, and Srinivasan (2014), and Jiang and Huang (2016) were applied to the TEP simulated benchmark. Jiang, Yan, and Huang (2016) applied their propositions to the TEP to compare results with Ghosh, Ramteke, and Srinivasan (2014), but also verified their method's performance in a real dataset from an oil industry. An automotive manufacturing dataset was the case studied by González and Sánchez (2010) and a real cigarette production was analyzed by Zhao and Gao (2017). Giannetti et al. (2014) and Jeong et al. (2012) developed their methods to deal particularly with complex, and very specific, manufacturing settings, such as foundry environment and molten carbonate fuel cell power plant, respectively. Finally, Wang and Tsung (2009) tested their method on simulated data.

2.5.3.1 Fault Detection

Works in Clusters 6 and 9 focused on improvements in fault detection. Five out of the seven methods in those clusters adapted PCA-based MSPC strategies.

The three methods in Cluster 6 used a preprocessing filter approach: one proposed a framework for batch process monitoring; the other two presented new CCs developed to monitor continuous processes. The performance of the framework proposed by Chu, Qin and Han (2004) was compared with results obtained from a traditional PCA-based fault detection method; in opposition to the later, a zero error rate in the detection of faults was verified using the framework. A drawback is that Chu, Qin and Han (2004)'s proposition operates with a large pre-specified number of normal and faulty process observations to establish decision boundaries between classes, demanding a large training dataset containing all possible process conditions. The framework was developed for batch process monitoring, but may also be applied to continuous processes. The ADR-2 CC (WANG; TSUNG, 2009) is able to switch automatically between projected statistics, choosing the most efficient to be used at each process step; applying the principle of dimension reduction guarantees the optimality of each statistic.

In simulated tests, the ADR scheme substantially improved both large and small shift detection requiring less computational power. Aiming at dealing with nonlinear multimode continuous processes, the TDB method (GE; GAO; SONG, 2011) and multi-model kernel PCA (MKPCA) showed much better monitoring performance than multi-model PCA, since TDB and MKPCA can handle the nonlinear data behavior in each operation mode. TDB stands out since its computational complexity is much lower than MKPCA. However, some aspects of TDB methods should be improved, such as the determination of the number of variables in each linear subspace, the assumption that the number of linear subspaces is the same in different operation modes, and the fact that there may be some situations under which the linear correlation is weak in every linear subspace degrading modeling performance.

Approaches in Cluster 9 applied a wrapper approach to MSPC in datasets from continuous and batch processes to promote a better fault detection when monitoring IC variables. There are four methods in the cluster (GHOSH; RAMTEKE; SRINIVASAN, 2014; GONZÁLEZ; SÁNCHEZ, 2010; GOURVÉNEC; CAPRON; MASSART, 2004; JEONG et al., 2012). One of the major objectives of the wrapper VS approach is to improve the accuracy of the methods. The integration of OPA and GA (GOURVÉNEC; CAPRON; MASSART, 2004) obtained reduced NIR spectra which was expected to promote better monitoring of a batch process. However, this was the only case in which there was no clear evidence of the improvement when the new method was compared to traditional OPA. One point of discussion is that the reduction in time promoted by the VS did not compensate the increase in computational cost due to the use of the wrapper approach. In search of a better T^2 chart performance, González and Sánchez (2010) selected and monitored only a subset of dominant variables. The T^2 charts constructed with the selected variables were more effective in the detection of simulated alarms than the chart that monitored all the original variables. The other two methods that integrate Cluster 9 aimed at overcoming limitations of PCA monitoring. In the first one, FA was used recursively to identify groups of variables to be monitored by PCA (JEONG et al., 2012). Type I and type II errors were reduced by more than half, and the total explained variance was increased when the method was compared to traditional PCA. In the other case, a reduced PCA model based on an optimal subset of variables from the training dataset (GHOSH; RAMTEKE; SRINIVASAN, 2014) was compared to the full PCA model, resulting in the reduction of both FAR and MDR when tested on validation data.

The detection delay was also shorter when the new method was applied, and a faster and more sensitive detection of multiple faults was achieved.

2.5.3.2 Fault Isolation

Methods in Clusters 7 and 10 share the objective of improving the monitoring of IC variables, being focused on the fault isolation task.

Cluster 7 comprises 2 methods that use Filter preprocessing VS approaches to adapt PCA-based MSPC methods. The BSPCA CC in Ge, Zhang, and Song (2010) was proposed for nonlinear process monitoring, and compared with the traditional PCA-based MSPC in a numerical example. Both methods successfully detected faults; however, using BSPCA it was possible to determine the subspace (PCS or RS) most responsible for the alarm. Ramp change faults, which are hardly detected using PCA-MSPC, were well identified through BSPCA. Fault isolation was efficiently performed using a reconstruction-based contribution plot method, both in the combined subspace and in each subspace separately. When compared to conventional PCA and kernel PCA for fault isolation in a simulated chemical industry data, BSPCA outperformed both methods in most fault cases, showing its feasibility and efficiency. The framework of the improved 2-D-DPCA modeling method (YAO et al., 2009) requires no a priori process knowledge, presenting a good potential to be applied in different batch processes. In a simulation study, 2-D-DPCA models with auto-determined ROS presented better performance, both for fault detection and isolation, when compared to 2-D-DPCA models with quarter-plane ROS.

Fault isolation using a wrapper VS approach that integrated GA and LASSO with MSPC methods was the proposition in the four methods assigned to Cluster 10.

The FBPCA (JIANG; YAN; HUANG, 2016) extended the method in Ghosh, Ramteke, and Srinivasan (2014) to include a fault isolation step. The performance of this new method was considered superior in most cases when compared to PCA and several PCA-based methods. The new method was also able to effectively detect faults as early as at the beginning of fault occurrence in a real dataset. That was verified through the low values of FAR and MDR obtained. Fault isolation was successfully achieved by the new FBPCA contribution plots, which more clearly separated responsible from non-responsible variables. Jiang and Huang (2016) also integrated GA with PCA in a distributed process monitoring framework. Monitoring results were evaluated using the same numerical example in Jiang, Yan and Huang (2016). The proposed framework was compared to global PCA, reduced PCA with one block (similar to Ghosh, Ramteke, and Srinivasan, 2014), distributed PCA with two blocks, and distributed PCA with three blocks (similar to Jiang, Yan and Huang, 2016). As the number of sub-blocks increases, there is a significant reduction on the number of non-detected fault points. Regarding fault isolation, as the fault magnitude increases, fault status can be successfully identified in general.

Chiang, Pell, and Seasholtz (2004)'s approach, which incorporated GA to FDA, was compared with T^2 and SPE statistic contribution charts for fault isolation in a simulated industrial process. The authors' method provides a more direct indication of the variables responsible for the fault. As process faults propagate to the majority of process variables, GA/FDA provided better consistency in identifying the faulty variables when compared to contribution charts.

The SDISSIM method proposed by Zhao and Gao (2017) integrates LASSO regression and the DISSIM method, and is used to isolate incipient faulty variables responsible for distortions in the underlying process covariance structure. The number of selected faulty variables and the missing reconstruction ratio (MRR; i.e. the ratio between alarms that have not been eliminated after removal of selected variables and the total number of alarms) were used as performance indicators. Applying SDISSIM, all the incipient faulty variables were correctly isolated resulting in the smallest MRR value, with all alarms eliminated after the removal of selected variables. In comparison, when reconstruction-based contribution and DISSIM-based methods were applied, more variables were wrongly isolated as faulty ones.

2.5.3.3 Fault Diagnosis

Closing our proposed categorization of articles, fault diagnosis was addressed by works in Cluster 8. The improvement of the fault diagnosis method of co-linearity index and penalty matrices, which used a Filter Postprocessing VS approach, allowed the evaluation of a dataset comprised of categorical and continuous variables. The definition of optimal process settings, which would assist engineers in the analysis of root causes, is possible since the method displays the noise free correlations between heterogeneous process variables and responses. Furthermore, the proposed data pre-treatment transformations were more robust to the presence of outliers and variables with skewed distributions.

2.6 CONCLUSION AND OPEN ISSUES

The growing dimensionality of datasets from industrial processes calls for adaptations on traditional MSPC methods. This systematic review presented the current state-of-the-art of VS methods integrated to MSPC, and answered three research questions. Limitations present in MSPC were associated with three main objectives that guided the development of the 30 methods reviewed here. Adaptations in projection methods such as PCA and PLS were responsible for the main improvements in MSPC, with LASSO regression, GA, and FS being the main VS methods applied. Fault isolation and detection were the main steps investigated in process monitoring. The 30 methods in this review were classified according the VS approach applied to integrate VS in MSPC, categorized according to objectives that guided MSPC improvement and the step of process monitoring they were aimed at, resulting in ten clusters of works. Methods covered in this review were published between 2002 and 2017, testifying the increasing attention given to the topic in the SPC literature.

Open Issues for future research

From the analysis of investigated methods five groups of research opportunities were identified. They provide an answer to our fourth research question (*“which research opportunities arise from gaps in the current state-of-the-art on the subject?”*), and are described next.

i) **New combinations of VS and MSPC methods.** Of the 27 quadrants in Figure 2.3 corresponding to combinations of VS approaches and MSPC objectives in different steps of process monitoring, only 10 are currently explored in the literature. That leaves several situations open to investigation. Examples include (*i*) enhancement of methods to exclusively monitor potential OOC variables through Wrapper approach aiming at better detecting and diagnosing faults, (*ii*) improvements in the monitoring of IC variables using Filter Postprocessing to better detect and isolate faults, (*iii*) development of new methods to better explain and predict response variables aiming at fault isolation using all VS approaches available, and (*iv*) development of methods to promote fault diagnosis using Filter Preprocessing and Wrapper approaches to achieve all objectives of MSPC adaptations (i.e. exclusive monitoring of potential OOC variables, better modeling and prediction of response variables, and improvement in the monitoring of IC variables).

ii) **Enhancements on existing methods.** Further developments on works presented in this review are suggested; for example: (*i*) use of different VS methods to improve MSPC (JIANG; WANG; TSUNG, 2012; NISHIMURA; MATSUURA; SUZUKI, 2015; SHINOZAKI; IIDA, 2017; WANG; JIANG, 2009), (*ii*) enhancement of MSPC methods not explored by authors in their original works (NISHIMURA; MATSUURA; SUZUKI, 2015; YAN; YAO, 2015), (*iii*) adaptation of methods to handle *nonnormal* data (GONZÁLEZ; SÁNCHEZ, 2010), and (*iv*) identification of a VS procedure to detect shifts in process dispersion (CAPIZZI; MASAROTTO, 2011). As methods were adapted to specific types of processes (mainly chemical, semiconductor and metallurgical industries), applying those methods to datasets originated from different industrial segments may confirm their robustness (ZARZO; FERRER, 2004).

iii) **Process monitoring in SPC.** Most methods reviewed in this paper focused fault detection or isolation. That points to research opportunities in the development of fault diagnosis methods, which were the subject of only two methods in this review.

iv) **Monitoring of batch processes.** Only 23% of the methods covered in this review discussed improvements in batch process monitoring. As this type of process is very frequent in industry (e.g. food, chemical, and pharmaceutical sectors), monitoring and optimizing its performance through VS appears as a promising research topic. As suggested by Anzanello and Fogliatto (2014), the insertion of a preliminary VS step in the analysis of n-way data arrays could be a starting point.

v) **Methods for phase I monitoring.** Methods presented in this review focused on improving the performance of Phase II of SPC. However, in some cases improvements in Phase I could lead to an easier monitoring of Phase II. In such context, Jiang, Wang, and Tsung (2012) discussed that the development of a VS chart for Phase I, and the development of a VS method for identifying variables responsible for OOC signals, are open issues to be studied.

2.7 REFERENCES

ABDELLA, G. M.; AL-KHALIFA, K. N.; KIM, S.; JEONG, M. K.; ELSAYED, E. A.; HAMOUDA, A. M. Variable Selection-based Multivariate Cumulative Sum Control Chart. **Quality and Reliability Engineering International**, v. 33, p. 565–578, 2017.

ANZANELLO, M. J.; FOGLIATTO, F. S. A review of recent variable selection methods in industrial and chemometrics applications. **European Journal of Industrial Engineering**, v. 8, n. 5, p. 619–645, 2014.

CAPIZZI, G. Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II. **Quality Engineering**, v. 27, n. May, p. to appear, 2015.

CAPIZZI, G.; MASAROTTO, G. A Least Angle Regression Control Chart for Multidimensional Data. **Technometrics**, v. 53, n. 3, p. 285–296, 2011.

CHIANG, L. H.; PELL, R. J.; SEASHOLTZ, M. B. Multivariate analysis of process data using robust statistical analysis and variable selection. **IFAC Proceedings Volumes**, v. 37, n. 1, p. 269–274, 2004.

CHU, Y.-H.; LEE, Y.-H.; HAN, C. Improved Quality Estimation and Knowledge Extraction in a Batch Process by Bootstrapping-Based Generalized Variable Selection. **Industrial & Engineering Chemistry Research**, v. 43, n. 11, p. 2680–2690, 2004.

CHU, Y.-H.; QIN, S. J.; HAN, C. Fault Detection and Operation Mode Identification Based on Pattern Classification with Variable Selection. **Industrial & Engineering Chemistry Research**, v. 43, p. 1701–1710, 2004.

GE, Z.; GAO, F.; SONG, Z. Two-dimensional Bayesian monitoring method for nonlinear multimode processes. **Chemical Engineering Science**, v. 66, p. 5173–5183, 2011.

GE, Z.; ZHANG, M.; SONG, Z. Nonlinear process monitoring based on linear subspace and Bayesian inference. **Journal of Process Control**, v. 20, p. 676–688, 2010.

GHOSH, K.; RAMTEKE, M.; SRINIVASAN, R. Optimal variable selection for effective statistical process monitoring. **Computers and Chemical Engineering**, v. 60, p. 260–276, 2014.

GIANNETTI, C.; RANSING, R. S.; RANSING, M. R.; BOULD, D. C.; GETHIN, D. T.; SIENZ, J. A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. **Computers & Industrial Engineering**, v. 72, p. 217–229, 2014.

GONZÁLEZ, I.; SÁNCHEZ, I. Variable Selection for Multivariate Statistical Process Control. **Journal of Quality Technology**, v. 42, n. 3, p. 242–259, 2010.

GOURVÉNEC, S.; LAMOTTE, C.; PESTIAUX, P.; MASSART, D. L. Use of the orthogonal projection approach (OPA) to monitor batch processes. **Applied Spectroscopy**, v. 57, n. 1, p. 80–87, 2003.

GOURVÉNEC, S.; CAPRON, X.; MASSART, D. L. Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection. **Analytica Chimica**

Acta, v. 519, p. 11–21, 2004.

JEONG, H.; CHO, S.; KIM, D.; PYUN, H.; HA, D.; HAN, C.; KANG, M.; JEONG, M.; LEE, S.A heuristic method of variable selection based on principal component analysis and factor analysis for monitoring in a 300 kW MCFC power plant. **International Journal of Hydrogen Energy**, v. 37, n. 15, p. 11394–11400, 2012.

JIANG, Q.; HUANG, B. Distributed monitoring for large-scale processes based on multivariate statistical analysis and Bayesian method. **Journal of Process Control**, v. 46, p. 75–83, 2016.

JIANG, Q; HUANG, B.; YAN, X. GMM and optimal principal components-based Bayesian method for multimode fault diagnos**is. Computers & Chemical Engineering**, v. 84, p. 338-349, 2016.

JIANG, Q.; YAN, X.; HUANG, B. Performance-Driven Distributed PCA Process Monitoring Based on Fault-Relevant Variable Selection and Bayesian Inference. **IEEE Transactions on Industrial Electronics**, v. 63, n. 1, p. 377–386, 2016.

JIANG, W.; WANG, K.; TSUNG, F. A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. **Journal of Quality Technology**, v. 44, n. 3, p. 209–230, 2012.

KANO, M.; HASEBE, S.; HASHIMOTO, L. Statistical process monitoring based on dissimilarity process data. **AlChe Journal**, v. 48, n. 6, p. 1231-1240, 2002.

KOURTI, T. Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. **International Journal of Adaptive Control and Signal Processing**, v. 19, p. 213–246, 2005.

KUANG, T. H.; YAN, Z.; YAO, Y. Multivariate fault isolation via variable selection in discriminant analysis. **Journal of Process Control**, v. 35, p. 30–40, 2015.

LI, W.; PU, X.; TSUNG, F.; XIANG, D. A robust self-starting spatial rank multivariate EWMA chart based on forward variable selection. **Computers & Industrial Engineering**, v. 103, p. 116–130, 2017.

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, v. 23, p. 35–44, 1999.

MEGAHED, F. M.; JONES-FARMER, A. A statistical process monitoring perspective on “big data”. In **Frontiers in Statistical Quality Control**. 11th ed., 2013. 21p.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SæBø, S. A review of variable selection methods in Partial Least Squares Regression. **Chemometrics and Intelligent Laboratory Systems**, v. 118, p. 62–69, 2012.

NISHIMURA, K.; MATSUURA, S.; SUZUKI, H. Multivariate EWMA control chart based on a variable selection using AIC for multivariate statistical process monitoring. **Statistics & Probability Letters**, v. 104, p. 7–13, 2015.

RANSING, R. S.; GIANNETTI, C.; RANSING, M. R.; JAMES, M. W. A coupled penalty matrix approach and principal component based co-linearity index technique to discover product specific foundry process knowledge from in-process data in order to reduce defects. **Computers in Industry**, v. 64, n. 5, p. 514–523, 2013.

SHINOZAKI, N.; IIDA, T. A variable selection method for detecting abnormality based on the T 2 test. **Communications in Statistics - Theory and Methods**, v. 46, n. 17, p. 8603–8617, 2017.

TUR, J. M. M. I.; CELLER, F.E.; HUBER, R. M; QIN, S. J. On the selection of variables for qualitative modelling of dynamical systems. **International Journal of General Systems**, v. 31, n. 5, p. 435–467, 2002.

VAN AELST, S.; WELSCH, R.; ZAMAR, R. H. Special Issue on Variable Selection and Robust Procedures. **Computational Statistics and Data Analysis**, v. 54, p. 2879–2882, 2010.

WANG, K.; JIANG, W. High-dimensional process monitoring and fault isolation via variable selection. **Journal of Quality Technology**, v. 41, n. 3, p. 247–258, 2009.

WANG, K.; TSUNG, F. An Adaptative Dimension Reduction Scheme for Monitoring Feedback-controlled Process. **Quality and Reliability Engineering International**, v. 25, p. 283–298, 2009.

WOODALL, W.; MONTGOMERY, D. Some current directions in the theory and application of statistical process monitoring. **Journal of Quality Technology**, v. 46, n. 1, p. 78–94, 2014.

YAN, Z.; KUANG, T. H.; YAO, Y. Multivariate fault isolation of batch processes via variable selection in partial least squares discriminant analysis. **ISA Transactions**, v. 70, p. 389–399, 2017.

YAN, Z.; YAO, Y. Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). **Chemometrics and Intelligent Laboratory Systems**, v. 146, p. 136–146, 2015.

YAO, Y.; DIAO, Y.; LU, N.; GAO, F. Two-dimensional dynamic principal component analysis with autodetermined support region. **Industrial and Engineering Chemistry Research**, v. 48, p. 837–843, 2009.

YUE, H. H.; QIN, S. J. Reconstruction-Based Fault Identification Using a Combined Index. **Industrial & Engineering Chemistry Research**, v. 40, n. 20, p. 4403–4414, 2001.

ZARZO, M.; FERRER, A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. **Chemometrics and Intelligent Laboratory Systems**, v. 73, n. 1, p. 15–27, 2004.

ZHAO, C.; GAO, F. A sparse dissimilarity analysis algorithm for incipient fault isolation with no priori fault information. **Control Engineering Practice**, v. 65, p. 70–82, 2017.

ZHAO, C.; SUN, Y.; GAO, F. A multiple-time-region (MTR)-based fault subspace decomposition and reconstruction modeling strategy for online fault diagnosis. **Industrial and Engineering Chemistry Research**, v. 51, n. 34, p. 11207–11217, 2012.

ZHAO, C.; WANG, W. Efficient faulty variable selection and parsimonious reconstruction modelling for fault isolation. **Journal of Process Control**, v. 38, p. 31–41, 2016.

ZOU, C.; NING, X.; TSUNG, F. LASSO-based multivariate linear profile monitoring. **Annals of Operations Research**, v. 192, p. 3–19, 2012.

ZOU, C.; QIU, P. Multivariate statistical process control using LASSO. **Journal of American Statistical Association**, v. 104, n. 488, p. 1586–1595, 2009.

3 ARTIGO 2 - STRATEGIES FOR SYNCHRONIZING CHOCOLATE CONCHING BATCH PROCESS DATA USING DYNAMIC TIME WARPING

Artigo submetido ao Journal of Food Science and Technology; em revisão.
Prêmio de Melhor Apresentação no 20th International Conference on Food Processing and Technology

Abstract

In batch processing, process control is typically carried out comparing trajectories of process variables with those in a reference set of batches that yielded products within specifications. However, one strong assumption of these schemes is that all batches have equal duration and are synchronized, which is often not satisfied in practice. To overcome that, dynamic time warping (DTW) methods may be used to synchronize stages and align the duration of batches. In this paper, three DTW methods are compared using supervised classification through the k-nearest neighbor technique to determine the reference set in a milk chocolate conching process. Four variables were monitored over time and a set of 62 batches with durations between 495 and 1,170 minutes was considered; 53% of the batches were known to be conforming based on lab test results and experts' evaluations. All three DTW methods were able to promote the alignment and synchronization of batches; however, the KMT method (KASSIDAS; MACGREGOR; TAYLOR, 1998) outperformed the others, presenting 93.7% accuracy, 97.2% sensitivity, and 90.3% specificity in batch classification as conforming and non-conforming. The drive current of the main motor was the most consistent variable from batch to batch, being deemed the most important to promote alignment and synchronization of the chocolate conching dataset.

Keywords: Batch process monitoring; Variable duration; Reference distribution; Dynamic Time Warping; Chocolate conching

3.1 INTRODUCTION

Quality management (QM) practices have become increasingly important in a large number of industrial sectors, being either motivated by rising consumers' expectations, government regulations, or increased market competition. Despite the benefits arising from using such practices, Dora et al. (2013) observed a limitation in the literature addressing QM targeted at requirements from the food sector. Analyzing a sample of micro, small and medium-sized food companies in Europe, the authors found that statistical process control (SPC) was used by only 15% of the companies. That agrees with Lim, Antony, and Albliwi (2014), who reported an incipient utilization of SPC in food companies when compared to other sectors.

In highly processed food products such as chocolate, the quality of raw materials used in the blend, the proportion in which they are used, and processing conditions are fundamental in establishing the quality of final products (CIDELL; ALBERTS, 2006). Chocolate conching is a high investment, time-consuming batch process carried out in a specific equipment which vigorously mixes components during 5 hours to 3 days under increasing temperature from 45 to 80°C in batches of 2 to 20 tons (BOLENZ; KUTSCHKE; LIPP, 2008; BOLENZ; MANSKE; LANGER, 2014; BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; DI MATTIA et al., 2014; FRANKE; TSCHEUSCHNER, 1991; OWUSU; PETERSEN; HEIMDAL, 2012; TORRES-MORENO et al., 2012). Conching is considered a key step in chocolate manufacturing, greatly affecting the product’s sensory performance and achievement of specific rheological properties (BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; BORDIN SCHUMACHER et al., 2009; DI MATTIA et al., 2014; OWUSU; PETERSEN; HEIMDAL, 2012). Chocolate acceptability (BOLENZ; MANSKE; LANGER, 2014; OWUSU; PETERSEN; HEIMDAL, 2012; PRAWIRA; BARRINGER, 2009) and the amount of retained cocoa antioxidants (DI MATTIA et al., 2014; GÜLTEKIN-ÖZGÜVEN; BERKTAS; ÖZÇELIK, 2016) are also affected by this process step.

During conching a large number of process variables are monitored; e.g. rotation speed of shovels, drive current of the main motor, and chocolate temperature (BÜHLER, 2010). Processing variations, particularly in batch temperature profile and duration, may lead to several quality problems in the final product; e.g. heterogeneous and unsmooth chocolate, absence of desirable flavors or presence of acid flavor, fatty mouthfeel, fat and sugar bloom, and abnormal viscosity, texture and chocolate flavor. Thus, time/temperature settings are important process parameters, acting as indicators of the chocolate’s final quality (CIDELL; ALBERTS, 2006; DI MATTIA et al., 2014; PRAWIRA; BARRINGER, 2009; TORRES-MORENO et al., 2012). However, process monitoring results are rarely used to predict the product’s final quality, which is often assessed after the product has been sent to the next processing step (KASSIDAS; MACGREGOR; TAYLOR, 1998). The use of principal component control charts to allow on-line and off-line monitoring of a massive amount of data emerging from multivariate batch processes, such as chocolate conching, has been extensively investigated (JACKSON; MUDHOLKAR, 1979; KOURTI; MACGREGOR, 1996; MACGREGOR, 1997; NOMIKOS; MACGREGOR, 1994, 1995). All schemes are based on the use of multiway principal components analysis (MPCA) to promote dimensionality reduction of process datasets. Such reduced data are then used to construct control charts that allow verifying the behavior of a new batch in comparison to a reference set of batches that yielded products within specifications.

A main assumption constrains the use of MPCA-based control charts in the monitoring of batch processes: all assessed batches must present the same duration, and be synchronized with respect to process stages (GONZÁLEZ-MARTÍNEZ; FERRER; WESTERHUIS, 2011; KASSIDAS; MACGREGOR; TAYLOR, 1998; RAMAKER et al., 2003). That is rarely the case in the conching step of chocolate processing. The typical time required to conch different types of chocolate may vary due to a number of factors; namely: (*i*) environmental, such as room temperature and humidity; (*ii*) operational, such as high feeding time of the conche, closed louvres during dry conching phase, and unplanned interruptions in processing; and (*iii*) quality-related, such as high moisture content of raw materials. These factors cause the duration of batches to vary, and limit the application of MPCA-based control charts in the conching process.

To overcome that, batch process datasets must be pretreated using methods that align and synchronize variables' trajectories (GARCÍA-MUÑOZ et al., 2003). One of such methods is the Dynamic Time Warping (DTW), which stretches, compresses and translates intervals of process variables' trajectories (GONZÁLEZ-MARTÍNEZ; WESTERHUIS; FERRER, 2013; WESTERHUIS; KOURTI; MACGREGOR, 1999). Gollmer and Posten (1996), for instance, used DTW to assess different process phases in a fermentation process using *Saccharomyces cerevisiae*. Similarly, Kassidas, MacGregor, and Taylor (1998) presented an iterative method based on DTW for the synchronization of batch trajectories from an industrial emulsion polymerization process before the application of Nomikos and MacGregor (1994, 1995)'s MPCA-based monitoring scheme. However, different DTW methods may lead to different sets of synchronized batches, and that may impact the performance of monitoring schemes, particularly with respect to the determination of the reference set of conforming batches against which future batches will be compared to draw conclusions about process stability, and predict the product's final quality.

The purpose of this study is to verify how different methods for alignment and synchronization affect the ability to correctly classify batches of a milk chocolate dataset as conforming and non-conforming. The majority of works on batch process monitoring focuses on process monitoring phase (phase 2 of SPC), in which control charts' parameters are already available from the analysis of reference batches (KOURTI, 2003; KOURTI; MACGREGOR, 1996; MACGREGOR, 1997; NOMIKOS; MACGREGOR, 1995). This paper focuses on the development of the monitoring in-control model (phase 1 of SPC), analyzing the impact that different methods for alignment and synchronization have on the determination of the reference set of conforming batches. To the best of our knowledge, no previous work addressed this problem.

3.2 MATERIALS AND METHODS

3.2.1 Chocolate conching dataset

Data from 69 conching batches of milk chocolate processed between April 2014 and January 2015 were collected from 2 Frisse conches type ELK (Bühler AG, Uzwill, Switzerland) in a chocolate manufacturing plant located in the south of Brazil. The expected batch duration for this type of chocolate in the plant is 495 min, but the sample displayed batches that took up to 1,170 min of processing for completion. The dataset included conforming and non-conforming batches, which were classified according to off-line analyses of viscosity and yield value. Fineness and moisture content were also evaluated by the Quality Control department. Results of measurements were obtained 5 minutes (fineness) and 15 minutes (viscosity and yield value) after sample collection, allowing corrections before the batch was sent to the moulding lines. Conching duration above the expected value usually occurs when rheological specifications are not achieved, requiring corrections in the chocolate mass. Moisture measurements were not considered for batch classification, since results were not available before 2 days after sample collection (note that abnormal moisture results point to raw materials out of specification or water leakage in the conching equipment). Among the several variables observable during conching, process experts recommend monitoring four which are strongly related to two key quality-related aspects of the output, namely chocolate flavor, which is influenced by (*i*) chocolate mass temperature, and chocolate rheology, which is influenced by (*ii*) main motor drive current, (*iii*) shovel speed, and (*iv*) frequency of soybean lecithin dosage.

3.2.2 Pre-treatment of data

Real datasets bring a new layer of complexity to data analyses, usually requiring adjustments before statistical methods can be correctly implemented. In this section the adjustments carried out in the milk chocolate conching dataset are described.

3.2.2.1 Dataset unfolding

The complete dataset may be viewed as a three-way data matrix of dimension $(I \times J \times K)$, where J are the process variables measured in K time intervals in I batches. The concept of batch-wise unfolding described by Kourti (2003) was adopted here, resulting in a two-dimensional matrix $(I \times JK)$, where each row corresponds to a I batch and each column corresponds to a JK unfolding variable. As described by Nomikos and MacGregor (1995), batch-wise unfolding is recommended when the objective is batch process monitoring with SPC methods.

3.2.2.2 Missing values

Variables were measured every 60 seconds as batches progressed. Missing observations occurred when sensors started registering data at different moments or when measurements were not taken due to sensor failure. Missing values were treated as follows: (*i*) when a missing observation (or sequence of missing observations) displayed equal values in the immediately preceding and subsequent entries, missing observations were replaced by that value; (*ii*) when a missing observation displayed different values in the immediately preceding and subsequent entries, it was replaced by the average of those values; and (*iii*) when a sequence of missing observations displayed different values in the immediately preceding and subsequent entries, missing observations were replaced by values that gradually increased/decreased, obtained by linear interpolation.

The largest sequence of missing observations found in the dataset corresponded to 3 minutes.

3.2.2.3 Outliers

From the set of 69 batches, 7 were excluded for presenting drive current and speed readings larger than the maximum admissable values informed by the conching equipment manufacturer. The outlier batches were identified plotting the original data as a function of time (data not shown) and usually occured before or after a missing value, indicating that it was the result of sensor failure.

3.2.3 Alignment and synchronization of multivariate batch datasets

Three alignment and synchronization methods were applied to the milk chocolate conching dataset. They were all comprised of a symmetric DTW algorithm followed by an asymmetric synchronization procedure. The three methods were proposed by Kassidas, MacGregor, and Taylor (1998), Ramaker et al. (2003), and González-Martínez, Ferrer, and Westerhuis (2011), respectively, and denoted hereafter as KMT, RSWS, and GFW.

The main principle of DTW is the non-linear warping of pairs of trajectories such that similar events are aligned and the best path through a grid of vector-to-vector distances is found, minimizing the total distance between trajectories. A weighted quadratic distance is used, where $\mathbf{W}$ is a diagonal matrix of weights for each variable giving their relative importance for the next iteration (KASSIDAS; MACGREGOR; TAYLOR, 1998). In KMT the weights provide an indication of which process variables are consistent from batch to batch, and synchronization of batch trajectories relies on those variables. In RSWS matrix $\mathbf{W}$ indicates which variables contain more warping information, explaining main occurrences in the process. It is clear that KMT and RSWS are grounded on different interpretations of weights and may lead to different warped datasets. Finally, GFW pursues a balance between the KMT and RSWS methods, defining a $\mathbf{W}$ matrix in which higher weights are assigned to variables that are consistent from batch to batch, while simultaneously providing meaningful warping information.

All batches (conforming and non-conforming) are aligned and synchronized with respect to a reference batch ($\boldsymbol{b_{ref}}$) such that all batches display the duration of

$\boldsymbol{b}_{\boldsymbol{ref}}$. The choice of $\boldsymbol{b}_{\boldsymbol{ref}}$ varies in each DTW method. In KMT and RSWS, $\boldsymbol{b}_{\boldsymbol{ref}}$ corresponds to the batch with duration closest to the average duration of conforming batches; in GFW $\boldsymbol{b}_{\boldsymbol{ref}}$ corresponds to the batch with duration closest to the median duration of conforming batches.

In this paper, two adaptations were implemented in the methods above. First, the band global constraint was not used. Such constraint aims at speeding up computational time by excluding certain regions of the dataset in which the optimal path may lie. This is particularly important in on-line applications; however, as analyses were conducted off-line the main goal is to find the optimal path across all batches and computational time became a minor issue. Second, a new convergence criterion for the weight matrix $\mathbf{W}$ is proposed. In this study, the iterative procedure was stopped when the difference between each one of the 4 variable weights in the current iteration and its corresponding ones in the immediately previous iteration was smaller than $|0.0010|$. Evaluation of results after weights achieved the convergence threshold showed that they were not significantly improved.

The alignment and synchronization methods tested were implemented in MATLAB R2012b; all codes are available upon request.

3.2.4 k-Nearest Neighbor classification method

The k-nearest neighbor (kNN) classification technique (BHATIA; VANDANA, 2010; COVER; HART, 1967; DUDA; HART; STORK, 2001) was used to classify the set of aligned and synchronized batches yielded by each DTW method in one of two categories: conforming or non-conforming. The k-nearest neighbors are considered those with the smallest Euclidean distance from the new observation. The treated dataset was partitioned such that 80% of the data were used as training portion, and 20% as test portion. Five different odd numbers of neighbors ($k =$1, 3, 5, 7, and 9) were tested and the best value of k was selected as the one that achieved the highest average classification accuracy in the test portion (in case of tie, the smallest k was chosen). Average accuracy was computed based on 100 iterations, being given by the number of correctly classified batches divided by the total number of classifications (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009). Sensitivity and specificity values were also computed for each value of k. Sensitivity corresponds to the *true* *positive rate* (i.e. the number of true conforming batches divided by the total number of batches classified as conforming), and specificity is the *true negative rate* (i.e. the number of true non-conforming batches divided by the total number of batches classified as non-conforming) (ALPAYDIN, 2010).

Since the objective is to refine the definition of the reference set of conforming batches to be used in phase 1 of SPC, the DTW method with largest values of classification accuracy and sensitivity was deemed best for the milk chocolate dataset under analysis. Specificity is also particularly relevant in situations where the focus is to minimize unnecessary production line interruptions and financial losses. The kNN was implemented in MATLAB R2012b; codes are available upon request.

3.3 RESULTS AND DISCUSSION

Three DTW methods were applied to promote the alignment and synchronization of batches in a milk chocolate conching dataset. Process experts selected 4 process variables to be monitored, with observations available at each minute. Variable *MS6* (‘Metering System #6’) shows the frequency of soybean lecithin dosage; variables *S* (‘Speed’) and *DC* (‘Drive Current’) give the rotation speed of the shovels and current of the main motor that promotes the rotation of the conche axis; variable *CT* (‘Chocolate Temperature’) gives the temperature of the mass during conching.

The dataset was pre-treated resulting in 62 batches classified as conforming (53%) and non-conforming (47%). Classification is an input of the DTW methods since the reference batch ($\boldsymbol{b_{ref}}$) is chosen from the subset of conforming batches. Pre-treated variables were scaled according to their average range, in all DTW methods, resulting in plots shown in Figure 3.1. Variable *MS6* displays occasional step changes in its level, being suitable to test the quality of synchronization. Variables *S*, *DC* and *CT*, although noisy, display an identifiable shape along the batch duration, proving that outliers were successfully removed. Differences in duration and lack of synchronization across batches may also be verified.

As previously mentioned, the choice of $\boldsymbol{b_{ref}}$ varied according to the DTW method being tested. The final length of batches treated by KMT and RSWS was 716 minutes, and the corresponding three-way warped data matrix had 177,568 data points.

In GFW the length of batches was warped to 699 minutes, and the corresponding three-way data matrix had 173,352 data points.

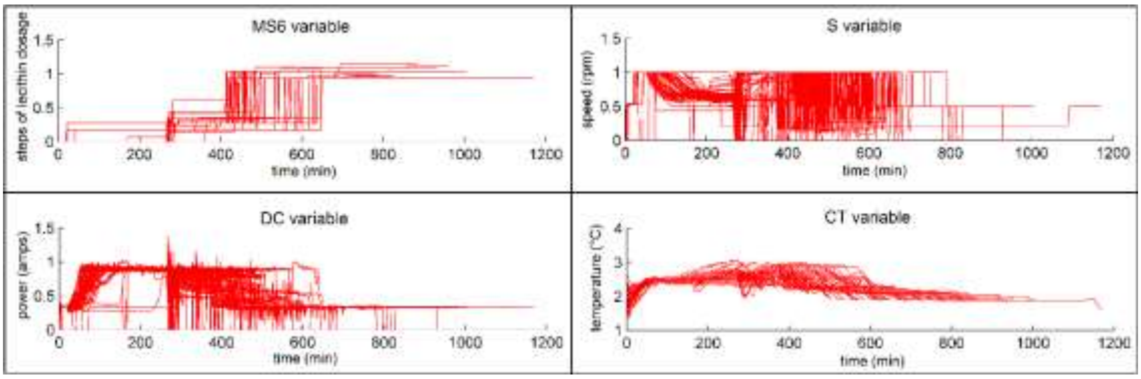


Figure 3. 1 Scaled pre-treated variables before synchronization

Convergence criteria varied according to the DTW method under test. The number of iterations required per method was 29 for the KMT, 34 for the GFW, and 36 for the RSWS. On average, 36 minutes were required to perform an iteration of the methods. Figure 3.2 shows the results of alignment and synchronization of batches obtained using each method. All methods were able to promote the synchronization of batches with similar results. Profiles of aligned and synchronized variable *MS6* obtained by KMT and RSWS were very similar, with small differences at the end of batches. The synchronization promoted by GFW differs from the other methods both in the duration and phase change points, which may be explained by the choice of $\boldsymbol{b_{ref}}$; its shape and synchronization results, however, are very similar to the other methods.

Profiles in Figure 3.2 allow identifying the progression of three industrial chocolate conching phases: feeding, main conching, and liquefaction (BÜHLER, 2010). In feeding phase, refined chocolate is loaded in the conche and temperature is raised up to 45°C. Emulsifiers may be added at this stage to confer a smoother structure to the chocolate mass, but they are often added at the liquefaction stage. Dry and plastic conching take place at main conching phase. In the dry conching, the chocolate mass is mixed under high temperature (70-80°C), and the louvres (upper windows of the conche) are maintained opened to allow evaporation of water and undesirable volatile acids, while aeration of product takes place (BOLENZ; KUTSCHKE; LIPP, 2008; BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; BÜHLER, 2010; PRAWIRA; BARRINGER, 2009). Transition to plastic conching starts when fat and/or soybean lecithin emulsifier are loaded (BOLENZ; THIESSENHUSEN; SCHÄPE, 2003;

GLICERINA et al., 2015). At this point, the current dependent drive changes from clockwise to anti-clockwise rotation in order to continue intense shearing and ensure better fat incorporation in the chocolate mass. Liquefaction phase starts when chocolate mass temperature is reduced to 45°C, and vanilla flavor and emulsifiers (soybean lecithin and/or polyglycerol polyricinoleate) are added while intense shear is applied. To prevent the mass from splashing out of the conche and to mix additives, louvres are closed and the rotation speed of shovels is reduced. Upon conclusion of this phase, a fluid chocolate mass should be available with rheological properties that ensure the efficiency of remaining process steps (such as pumping, tempering, and moulding) and sensory characteristics (such as consistency, smooth texture, melting, mouthfeel, and snap) (AFOAKWA et al., 2008; BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; BÜHLER, 2010; FRANKE; TSCHEUSCHNER, 1991; GLICERINA et al., 2013; OWUSU; PETERSEN; HEIMDAL, 2012; PRAWIRA; BARRINGER, 2009). To verify that, a sample of the mass is collected and its rheological parameters are assessed. If results are not in accordance with industrial standards, a new addition of emulsifiers or fat is usually required to correct chocolate mass' rheology.

From the profile of variable *MS6* two additions of soybean lecithin emulsifier are identifiable: at minutes 19 and 496 for KMT and RSWS, and at minutes 18 and 435 for GFW. The first addition takes place at the beginning of the feeding phase, and the second at the beginning of the liquefaction phase. Smaller step changes noticeable at the end of batches are related to lecithin additions aiming at correcting viscosity and yield values of non-conforming batches.

Analyzing the profile of variable *S*, several oscillations are observable at the beginning of batches until a maximum permissible speed plateau is reached, favoring an adequate refined mass distribution in the conche as feeding occurs. The end of the chocolate feeding phase takes place at minute 133 (for KMT and RSWS), after the speed is reduced to ensure an adequate mass aeration during dry conching phase. At minute 209 vegetable fat is added and the speed increases to allow plastic conching to occur, when kneading of the chocolate mass is required to better incorporate the fat. The power (variable *DC*) remains high and constant until the current dependent drive changes rotation from clockwise to anti-clockwise at minute 274 (for KMT and RSWS). Speed remains high and the power starts to decrease, as the chocolate mass needs progressively lower power to knead. When the mass is sufficiently liquid, speed and power decrease to lower the shear as the chocolate mass is now very soft and offers little resistance to the conching tools (such speed reduction also avoids the liquid mass to splash out of the equipment).

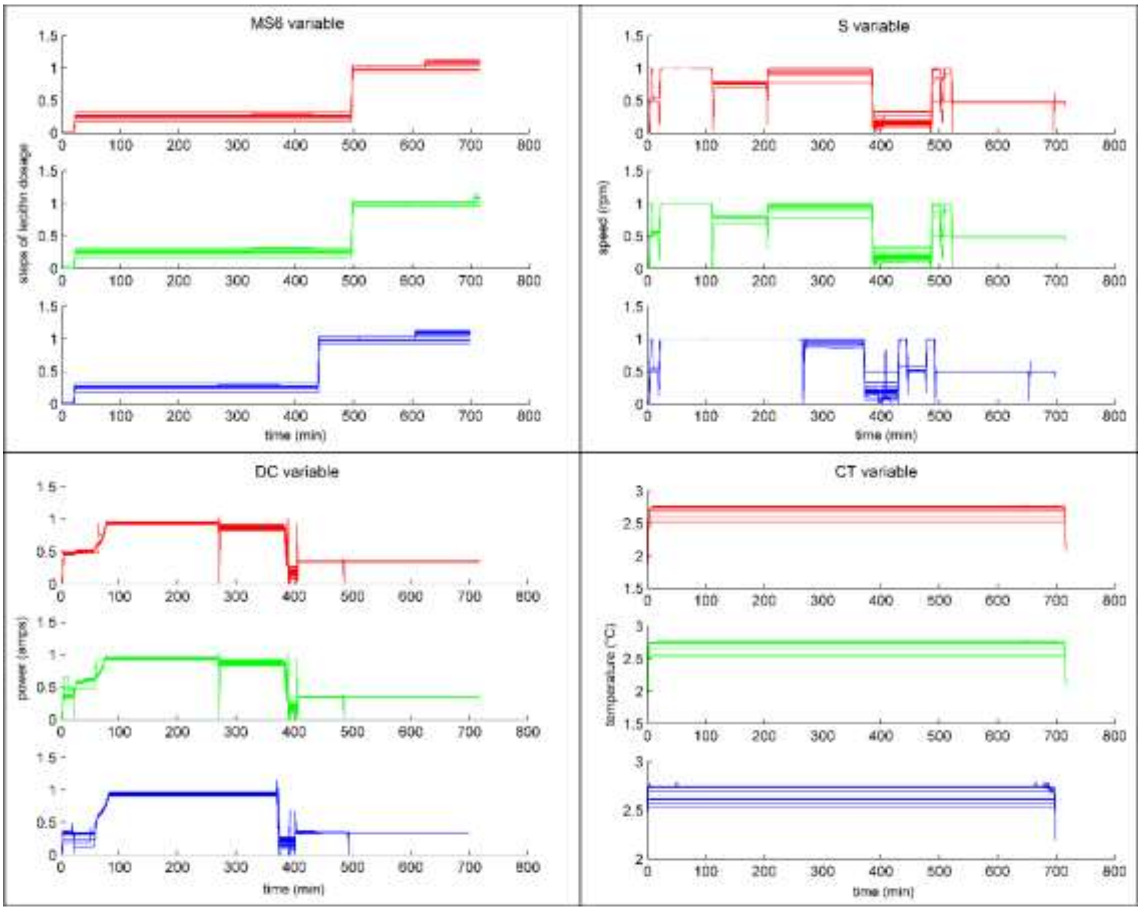


Figure 3. 2 Variables' profiles after synchronization using (a) KMT (top), (b) RSWS (middle), and (c) GFW (bottom) methods

Soybean lecithin is then added to the mass at minute 496, starting the liquefaction phase. Variable *S* displays a peak at this point since intense shear is required to properly mix emulsifiers into the chocolate mass. In the GFW method, the behavior of variables *S* and *DC* is not so obvious; however, rotation changeover and speed reduction at the end of the plastic conching are observable after minute 265 of processing. The temperature gets higher when the feeding phase ends and should reduce when the liquefaction phase starts; however, variable *CT* does not show clearly such temperature variations in any of the methods analyzed.

Figure 3.3 shows the progression in variables' weights as iterations of the DTW methods took place. At the start of the synchronization procedure it is not possible to average the batch trajectories since they have different durations, so **W** is assumed to be an identity matrix. From the fourth iteration on, the $\mathbf{W}$ matrix considered the average trajectory batch to be $\boldsymbol{b_{ref}}$. That explains the variation in weights in early iterations.

Depending on each method's objective in considering $\mathbf{W}$ in the warping process, different information is extracted from the synchronization.

Variable *DC* is the one with higher increment in synchronization capability if weights of the KMT method are analyzed, representing 52.19% of the total weight at the final iteration, and deemed the most consistent variable from batch to batch. This consistency is evident in variable *DC*'s profile in Figure 3.2 since its deviation from the average trajectory is almost non-existent. On the other hand, variable *DC* was not efficient in explaining the main occurrences in the process, representing only 3.45% of the total weight in the RSWS method. Variables *S* and *MS6* presented some dispersion between synchronized batches, being assigned lower weights than *DC* (22.05% and 20.21% respectively). Finally, variable *CT* was the most dispersed across batches (represented by a large variation band in Figure 3.2), being assigned the lowest weight in KMT method (5.55%).

In the RSWS method the largest weight is assigned to the variable with largest dispersion across batches. Thus, variable *CT* was the one with highest warping information, accounting for 82.95% of the total weight. All other variables displayed lower weight values, since their deviation from average trajectory was smaller. Variable *DC* displayed the lowest weight (3.45%).

In the GFW method weights were distributed in a narrower range: from 28.8% (variable *MS6*) to 20.12% (variable *CT*) due to a better balance between consistency and warping information promoted by the method.

Results in Figure 3.2 suggest that synchronization may be dependent on the choice of $\boldsymbol{b_{ref}}$. To check that, the outputs of all methods implemented with the same $\boldsymbol{b_{ref}}$ were evaluated (data not shown). First, KMT and RSWS were run using $\boldsymbol{b_{ref}}$ proposed in GFW (i.e. the batch with duration closest to the median duration of conforming batches); then, GFW was run using $\boldsymbol{b_{ref}}$ proposed in KMT and RSWS (i.e. the batch with duration closest to the average duration of conforming batches). Results showed that the choice of $\boldsymbol{b_{ref}}$ defines the length and phase changing points of synchronized trajectories, but does not impact on the quality of synchronization or the importance of variables in $\mathbf{W}$ matrix (the most consistent variable in KMT and the variable with the most warping information in RSWS remains the same, regardless of

$\boldsymbol{b_{ref}}$). Batches synchronized with the shorter trajectory length (699 minutes) required more iterations to convergence than batches synchronized with the longer trajectory length (716 minutes). That suggests that number of iterations is not directly related to length, but to shape and dispersion of the chosen $\boldsymbol{b_{ref}}$.

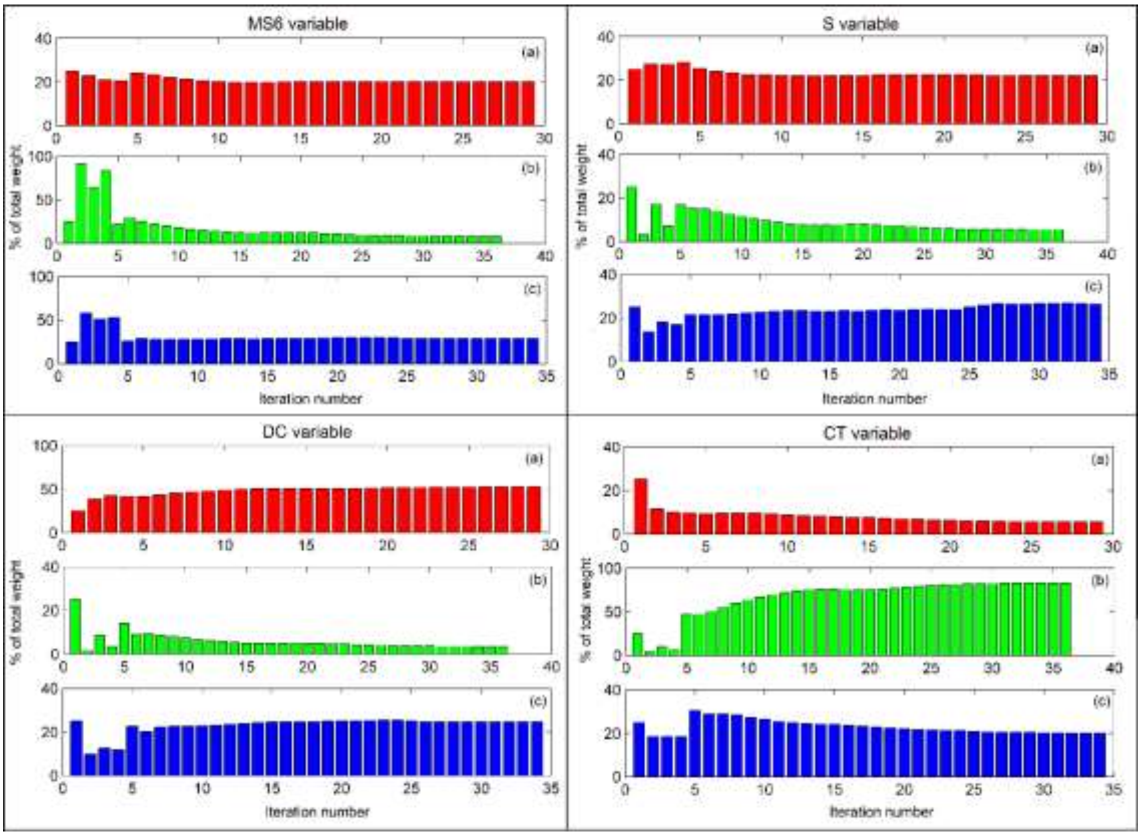


Figure 3. 3 Percentage of total weight per iteration number by (a) KMT (top), (b) RSWS (middle), and (c) GFW methods (bottom)

In addition to evaluating synchronization performance, the goal is to verify how each DTW method affected the classification of synchronized batches in conforming and non-conforming, using a supervised classification technique. The kNN classification technique was chosen for that. First, the dataset of synchronized batches was split into training and test portions, and the influence of the number of iterations in the final classification was evaluated. When performing 100, 200, 300, 400 and 500 iterations of the kNN the resulting standard deviation of average accuracy, sensitivity and specificity was lower than 0.7%, and the difference in performance as a function of the number of iterations was deemed not significant. Therefore, to minimize computational effort, analyses reported next resulted from running 100 iterations of the algorithm.

Running the kNN algorithm in the test portion of the chocolate conching dataset, varying parameter k from 1 to 9, the values of average and standard deviation for three classification performance metrics were obtained: accuracy, sensitivity, and specificity. Results are given in Table 3.1. In the conching process, accuracy of classification is of major importance: correct classification of batches is key to an adequate SPC. A higher specificity value is also crucial to avoid that non-conforming batches are misclassified as conforming, and released to the moulding lines. As batches are classified according to rheological parameters, a low specificity potentially leads to a number of technical problems in the process, such as imperfect covering of bonbons, poor spreading of chocolate mass in bar molds, and blocking of pipes and moving parts of the moulding equipment requiring interruptions in the production line for cleaning and disposal of products. That results in time and financial losses to the manufacturer. On the other hand, high sensitivity values are desired to better define the reference set of conforming batches to be used in the development of the monitoring in-control SPC model. The KMT method with $k = 3$ was selected as the one with the best combination of classification performance metrics. It yielded the highest average accuracy across methods and the second higher sensitivity and specificity average values. The RSWS method attained good results for $k = 1$, however with lower accuracy, sensitivity and specificity average results than the KMT method. Analyzing these results, it is possible to conclude that the choice of DTW method influence the ability of classifying batches in conforming or non-conforming.

RSWS' specificity with $k = 1$ resulted 2% lower than that of the KMT method with $k = 3$. Considering the number of batches processed in the plant analyzed, that corresponds to the misclassification of 2 batches per year; i.e. using the RSWS method to obtain the reference distribution to be used in SPC would imply in 12 tons of non-conforming chocolate mass being sent to the production line, leading to a series of technical problems. Losses will be greater proportionally to the deviation of the chocolate mass viscosity from its target value. Considering that the milk chocolate mass is used in the production of bonbons, and that non-conforming masses force the production line to halt, that would represent a total financial loss of US$ 1,700.00 in labor, and US$ 17,000.00 in raw materials.

Table 3. 1 Classification accuracy, sensitivity, and specificity for synchronized batches; all values given in percentage

		KMT method					RSWS method					GFW method				
		k					k					k				
		1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
Accuracy	Average	93.5	***93.7***	91.4	90.8	90.3	91.6	90.9	90.3	88.8	87.9	89.6	91.3	92.0	92.0	91.3
	Standard Deviation	6.9	6.5	8.0	8.6	8.7	6.5	7.3	7.1	8.6	8.9	7.9	7.1	6.9	7.3	7.7
Sensitivity	Average	96.0	***97.2***	95.8	94.9	93.2	95.4	97.0	95.3	93.3	91.4	90.6	95.4	96.6	97.8	97.9
	Standard Deviation	9.1	6.5	7.5	8.3	9.5	8.3	6.5	8.3	9.2	9.9	11.5	8.1	7.4	5.3	5.1
Specificity	Average	91.4	***90.3***	87.6	88.1	88.7	88.3	84.6	85.9	85.3	85.2	88.8	87.5	87.7	86.8	85.3
	Standard Deviation	11.6	12.0	14.2	14.5	15.0	13.1	14.8	13.5	15.2	15.4	12.0	12.6	12.5	13.3	15.0

In the assessed plant, 47% of the produced batches required some sort of correction. The average time to correct non-conforming masses was 52 minutes; assuming that production batches were expected to last average 495 minutes, 2 out of 18 planned batches were not produced due to such corrections. In addition, extra expenditure with raw materials required to correct non-conforming masses (e.g. soybean lecithin and vegetable fat) were also noteworthy, amounting to US$ 7,500.00 per year. The described losses in processing time and raw material can be substantially minimized by correctly implementing SPC in the conching equipment.

The kNN algorithm was also run to evaluate if different choices of $\boldsymbol{b}_{\boldsymbol{ref}}$ had an influence on the correct classification of batches (data not shown). When $\boldsymbol{b}_{\boldsymbol{ref}}$ was not the one suggested in the original DTW method, more k neighbors were required to obtain similar classification performance results. For example, when KMT used $\boldsymbol{b}_{\boldsymbol{ref}}$ suggested in the original method, 3 neighbors were required to obtain the highest accuracy (93.7%); on the other hand, when using $\boldsymbol{b}_{\boldsymbol{ref}}$ recommended in GFW 7 neighbors were required to obtain similar accuracy (93.3%). The same occurred with RSWS and GFW.

Correct classification of batches is key when determining the reference distribution of conforming batches to be used in SPC. With that in view, KMT was deemed the most adequate DTW method to promote alignment and synchronization of batches in the chocolate conching dataset: it was the method that least affected kNN's ability to correctly classify batches in conforming and non-conforming classes. Due to its consistency, variable *DC* was deemed the most important to promote the alignment and synchronization of the chocolate conching dataset.

3.4 CONCLUSION

In chocolate manufacturing process, batches typically do not present the same duration. As a consequence, traditional techniques as MPCA-based charts are not suitable for process control and monitoring. To address that issue, this paper compared three DTW methods that align and synchronize process variables' trajectories aimed at properly determining the reference distribution for multivariate statistical process control. Findings suggested the KMT method as the best DTW option for aligning and synchronizing a milk chocolate conching dataset. Such method was recommended due to the lowest number of iterations required to achieve convergence and highest average accuracy in the testing portion using the kNN classification technique. Future research includes the development of approaches focused on dimension reduction of datasets through variable selection.

3.5 REFERENCES

AFOAKWA, E. O.; PATERSON, A.; FOWLER M.; VIEIRA J. Relationship between rheological , textural and melting properties of dark chocolate as influenced by particle size distribution and composition. **European Food Research Technology**, v. 227, p. 1215–1223, 2008.

ALPAYDIN, E. **Introduction to machine learning.** 2ed. Cambridge: The MIT Press, 2010. 537p.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratory Systems**, v. 97, p. 111–117, 2009.

BHATIA, N.; VANDANA, A. Survey of Nearest Neighbor Techniques. International **Journal of Computer Science and Information Security**, v. 8, n. 2, p. 302–305, 2010.

BOLENZ, S.; KUTSCHKE, E.; LIPP, E. Using extra dry milk ingredients for accelerated conching of milk chocolate. **European Food Research Technology**, v. 227, p. 1677–1685, 2008.

BOLENZ, S.; MANSKE, A.; LANGER, M. Improvement of process parameters and evaluation of milk chocolates made by the new coarse conching process. **European Food Research Technology**, v. 238, p. 863–874, 2014.

BOLENZ, S.; THIESSENHUSEN, T.; SCHÄPE, R. Fast conching for milk chocolate.

European Food Research Technology, v. 218, p. 62–67, 2003.

BÜHLER, A. **Operating instructions ELK / DÜC conches**. Uzwill, Switzerland, 2010. 108p.

CIDELL, J. L.; ALBERTS, H. C. Constructing quality: The multinational histories of chocolate. **Geoforum**, v. 37, n. 6, p. 999–1007, 2006.

COVER, T. M.; HART, P. E. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

DORA M.; KUMAR M.; VAN GOUBERGEN D.; MOLNAR A., GELLYBCK X. Food quality management system: reviewing assessment strategies and a feasibility study for European food small and medium-sized enterprises. **Food Control**, v. 31, p 607–616, 2013.

DI MATTIA, C.; MARTUSCELLI M.; SACCHETTI, G.; BEHEYDT B.; MASTROCOLA D.; PITTIA P. Effect of different conching processes on procyanidin content and antioxidant properties of chocolate. **Food Research International**, v. 63, p. 367–372, 2014.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2ed. New York: John Wiley, 2001. 637p.

FRANKE, K.; TSCHEUSCHNER, H.-D. Modelling of the continuous high shear rate conching process for chocolate. **Joumal of Food Engineering**, v. 14, p. 103–115, 1991.

GARCÍA-MUÑOZ, S.; KOURTI, T; MACGREGOR, J. F.; MATEOS, A. G.; MURPHY G. Troubleshooting of an industrial batch process using multivariate methods. **Industrial Engineering Chemistry Research**, v. 42, p. 3592–3601, 2003.

GLICERINA, V.; BALESTRA, F.; DALLA ROSA M.; ROMANI, S. Rheological , textural and calorimetric modifications of dark chocolate during process. **Journal of Food Engineering**, v. 119, p. 173–179, 2013.

GLICERINA, V.; BALESTRA, F.; DALLA ROSA M.; ROMANI, S. Effect of manufacturing process on the microstructural and rheological properties of milk chocolate. J**OURNAL OF FOOD ENGINEERING**, v. 145, p. 45–50, 2015.

GOLLMER, K; POSTEN, C. Supervision of bioprocesses using a dynamic time warping algorithm. **Control Engineering Practice**, v. 4, n. 9, p. 1287–1295, 1996.

GONZÁLEZ-MARTÍNEZ, J. M.; FERRER, A.; WESTERHUIS, J. A. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. **Chemometrics and Intelligent Laboratory Systems**, v. 105, p. 195–206, 2011.

GONZÁLEZ-MARTÍNEZ, J. M.; WESTERHUIS, J. A.; FERRER, A. Using warping information for batch process monitoring and fault classification. **Chemometrics and Intelligent Laboratory Systems**, v. 127, p. 210–217, 2013.

GÜLTEKIN-ÖZGÜVEN, M.; BERKTAS, I.; ÖZÇELIK, B. Influence of processing conditions on procyanidin profiles and antioxidant capacity of chocolates : Optimization of dark chocolate manufacturing by response surface methodology. **Food Science and**

Technology, v. 66, p. 252–259, 2016.

JACKSON, J. E.; MUDHOLKAR, G. S. Control procedures for residuals associated with principal component analysis. **Technometrics**, v. 21, n. 3, p. 341-349, 1979.

KASSIDAS, A.; MACGREGOR, J. F.; TAYLOR, P. A. Synchronization of batch trajectories using dynamic time warping. **AIChE Journal**, v. 44, n. 4, p. 864–875, 1998.

KOURTI, T.; MACGREGOR, J. F. Multivariate SPC Methods for Process and Product Monitoring. **Journal of Quality Technology**, v. 28, p. 409-428, 1996.

KOURTI, T. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. **Annual Reviews in Control**, v. 27, p. 131–139, 2003.

LIM, S. A. H.; ANTONY J.; ALBLIWI S. Statistical process control (SPC) in the food industry – a systematic review and future research agenda. **Trends in Food Science and Technology**, v. 37, p. 137–151, 2014.

MACGREGOR, J. F. Using on-line process data to improve quality: challenges for statisticians. **International Statistical Review**, v. 65, p. 309–323, 1997.

NOMIKOS, P.; MACGREGOR, J. F. Monitoring batch processes using multiway principal component analysis. **AlChE Journal**, v. 40, n. 8, p. 1361–1375, 1994.

NOMIKOS, P.; MACGREGOR, J. F. Multivariate SPC charts for monitoring batch processes. **Technometrics**, v. 37, n. 1, p. 41-59, 1995.

OWUSU, M.; PETERSEN, M. A.; HEIMDAL, H. Effect of fermentation method, roasting and conching conditions on the roma of dark chocolate. **Journal of Food Processing and Preservation**, v. 36, p. 446–456, 2012.

PRAWIRA, M.; BARRINGER, S. A. Effects of conching time and ingredients on preference of milk chocolate. **Journal of Food Processing and Preservation**, v. 33, p. 571–589, 2009.

RAMAKER, H. J.; VAN SPRANG, E. N. M.; WESTERHUIS, J. A.; SMILDE, A. K. Dynamic time warping of spectroscopic batch data. **Analytica Chimica Acta**, v. 498, p. 133–153, 2003.

SCHUMACHER, A. B.; BRANDELLI, A.; SCHUMACHER, E. W.; MACEDO F. C.; PIETA L.; KLUG, T. V.; DE JONG E. V. Development and evaluation of a laboratory scale conch for chocolate production. **International Journal of Food Science and Technology**, v. 44, p. 616–622, 2009.

TORRES-MORENO, M; TARREGA, A.; COSTELL, E.; BLANCH C. Dark chocolate acceptability : influence of cocoa origin and processing conditions. **Journal of Science of Food Agriculture**, v. 92, p. 404–411, 2012.

WESTERHUIS, J. A.; KOURTI, T.; MACGREGOR, J. F. Comparing alternative approaches for multivariate statistical analysis of bacth process data. **Jornal of Chemometrics**, v. 13, p. 397–413, 1999.

4 ARTIGO 3 – FAULT DETECTION IN BATCH PROCESSES THROUGH VARIABLE SELECTION INTEGRATED TO MULTIWAY PRINCIPAL COMPONENT ANALYSIS

Abstract

The main purpose of fault detection in batch process monitoring is to identify batches displaying atypical behavior in comparison to normal operating data. The current growth in the number of measurable variables due to process automation yields datasets in which the number of variables is much larger than the number of batches. That may compromise the performance of Multiway Principal Component Analysis (MPCA), which is the most popular quality control approach used in batch processes. To overcome that, new strategies to handle high-dimensional datasets become necessary. In this paper we propose the Pareto Variable Selection (PVS) – MPCA method to monitor batch processes described by high-dimensional datasets. The main idea of PVS-MPCA is to select process variables that promote the best classification of production batches in conforming or non-conforming, prior to the construction of T^2 and Q control charts used to monitor batch performance. Our proposition was applied to a real dataset from a chocolate conching batch operation and compared to classical MPCA-based monitoring. PVS-MPCA promoted a reduction of 85.18% in false alarm rate retaining only 5 unfolded variables, in opposition to 2,864 unfolded variables used in classical MPCA. The missed detection rate was null, ensuring that only conforming batches were released to the production line.

Keywords: Batch process; Variable selection; Fault detection; Multiway Principal Component Analysis; High-dimensional data

4.1 INTRODUCTION

Batch processes are characterized as having finite duration, unsteady state, and non-linear nature (NOMIKOS; MACGREGOR, 1995a, 1995b). Generic steps of a batch process are: (*i*) charging batch contents, (*ii*) processing contents according to a protocol, and (*iii*) discharging processed contents. Even though batch processes are described by a large number of informative process variables, output quality is typically defined by comparing results of laboratory analyses with a predefined quality standard (KOSANOVICH; DAHL; PIOVOSO, 1996; NOMIKOS; MACGREGOR, 1995b).

To use process information in the definition of batch output quality, Nomikos and MacGregor (1994; 1995b) proposed calculating fault detection statistics in a reduced space obtained through multiway principal component analysis (MPCA) and monitoring their behavior using multivariate control charts. Chart parameters are determined using a reference distribution model obtained mining a historical database of past batches that yielded conforming outputs. The goal of fault detection through MPCA monitoring is to identify correctly batches displaying atypical behaviors when compared to normal operating data. Such potentially non-conforming batches are then investigated to identify process variables responsible for process failure, in a fault isolation step (WOODALL; MONTGOMERY, 2014; ZHAO; WANG, 2016). To attain successful batch process quality monitoring it is key to correctly classify batches as conforming or non-conforming.

The wide spread of sensor networks and distributed control systems in industry gave rise to a myriad of high dimensional datasets, comprised of measurements taken on hundreds of process variables during batch progression (JIANG; HUANG, 2016; WOODALL; MONTGOMERY, 2014). In classical PCA theory the asymptotic normality of eigenvalues and eigenvectors is established under a fixed model with dimension J (variables), as the number of I (observations or batches) tends to infinity (WANG; FAN, 2017). If J increases in parallel with I, standard PCA yields consistent estimates of the principal eigenvectors if and only $J/I \rightarrow 0$. However, in many datasets obtained from industrial applications, which is the case of batch processes with JK unfolding variables, JK is comparable to, or larger than the sample size I, with $JK/I \rightarrow \infty$ (JOHNSTONE; LU, 2009; LEE; LEE; PARK, 2012). When that is the case PCA yields inconsistent results with the sample covariance matrix being a notoriously bad estimator with substantial different eigen-structure from the original population (AMINI, 2011; WANG; FAN, 2017). That compromises the feasibility of MPCA-based multivariate monitoring methods in the fault detection and isolation steps, calling for new approaches to handle high-dimensional sets obtained from industrial batch processes (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012; BISGAARD, 2012; JIANG; HUANG, 2016).

Ideally, variables responsible for fault effects are distinguished from others and only a few key process variables are analyzed, expediting the process of diagnosing special events and enhancing the understanding of the fault generation mechanism (WANG; JIANG, 2009; ZHAO; WANG, 2016). To accomplish that it is important to decide which variables should be included in the reference model such that fault effects are more effectively explored. Variable selection methods can help with that decision; their use, integrated to multivariate statistical process control (MSPC), has considerably increased in the past five years as reported by Capizzi (2015) and Peres and Fogliatto (2018). Normally, hundreds (or even thousands) of variables are used for fault detection in batch processes which could impair the correct classification of batches in conforming and non-conforming due to their complex correlation structure (YAN; KUANG; YAO, 2017). Thus, identifying the subset of variables able to provide a better batch classification leads to the recovering of PCA consistency (see JOHNSTONE; LU, 2009; LEE; LEE; PARK, 2012; LUO et al., 2017), and improved process monitoring.

In this paper, we propose the Pareto Variable Selection – Multiway Principal Component Analysis (PVS–MPCA) method for fault detection using control charts (CCs). Our method is able to handle high-dimensional datasets in which $JK \gg I$. PVS-MPCA starts by selecting the process variables that best discriminate production batches as conforming or non-conforming. For that, Partial Least Squares Discriminant Analysis (PLS-DA) is performed and its output parameters used to computed a variable importance index (VII). Then, a wrapper variable selection procedure using VII and the k-Nearest neighbor (kNN) classification technique is run. An initial subset comprised of all variables is split by 5-fold cross validation and kNN is used to classify batches and compute classification accuracy on the testing portion. Subsequently, the variable with the lowest VII is removed, batches are re-classified using remaining process variables and classification accuracy is computed again. Variable elimination is carried out recursively until a single variable was left. A number of candidate sets (CS) of selected variables is chosen; using the Pareto Optimality (PO) analysis the Pareto frontier sets are identified and the variable subset with the smallest Euclidean distance from an ideal solution is selected as the Pareto Variable Selection (PVS) subset. Hotelling's T^2 and Q residuals CCs are then built to define the reference distribution using only the PVS subset of selected variables in an iterative procedure that eliminates batches outside the in-control region. Batches considered conforming by Phase I CCs are then used to build the MPCA monitoring model. For offline fault detection new batches have their T^2 and Q statistics computed and are monitored using Phase II CCs.

Finally, monitoring performance of CCs is measured by the cumulative error rate, that combines false alarm (FAR) and missed detection (MDR) rates.

There are two contributions here. First, we propose an off-line fault detection method that incorporates a variable selection step prior to the development of MPCA-based control charts aimed at monitoring high dimensional batch process datasets. Second, the fault detection performance of the new method is compared to that of the standard MPCA-based CCs method, showing its potential to mitigate PCA limitations that arise in the analysis of datasets in which $JK \gg I$.

The rest of the paper is organized as follows. In the section 4.2 we give the background of MPCA fault detection and variable selection methods used in our proposition. The proposed PVS-MPCA method is presented in section 4.3 and applied to real data from a chocolate manufacturer in section 4.4. Finally, conclusions and future research are presented in section 4.5.

4.2 BACKGROUND

4.2.1 Fault Detection using Batch-wise Multiway Principal Component Analysis (MPCA)

Consider process data describing the progression of batches organized in a three-way matrix $\mathbf{X}\,(I \times J \times K)$, such that J variables are measured at K time intervals in a sample of I batches. Measured variables are likely to display autocorrelation in addition to being correlated with one another at any given time during batch progression (DE OLIVEIRA; MARCONDES FILHO, 2018; KOSANOVICH; DAHL; PIOVOSO, 1996; YAN; KUANG; YAO, 2017; ZHAOMIN; QINGCHAO; XUEFENG, 2014). Aiming at monitoring batch-to-batch variation MPCA promotes the batch-wise unfolding of matrix $\mathbf{X}$ (Figure 4.1) to obtain a two-dimensional data matrix $\mathbf{X}_B$ with dimension $(I \times JK)$ on which principal component analysis (PCA) is performed. Through the unfolding procedure it is possible to address both the auto and cross-correlation of variables within a batch, as well as the variability across batches (EMPARÁN et al., 2012; KHOSRAVI; MELÉNDEZ; COLOMER, 2009; KOURTI, 2003; NOMIKOS; MACGREGOR, 1994).

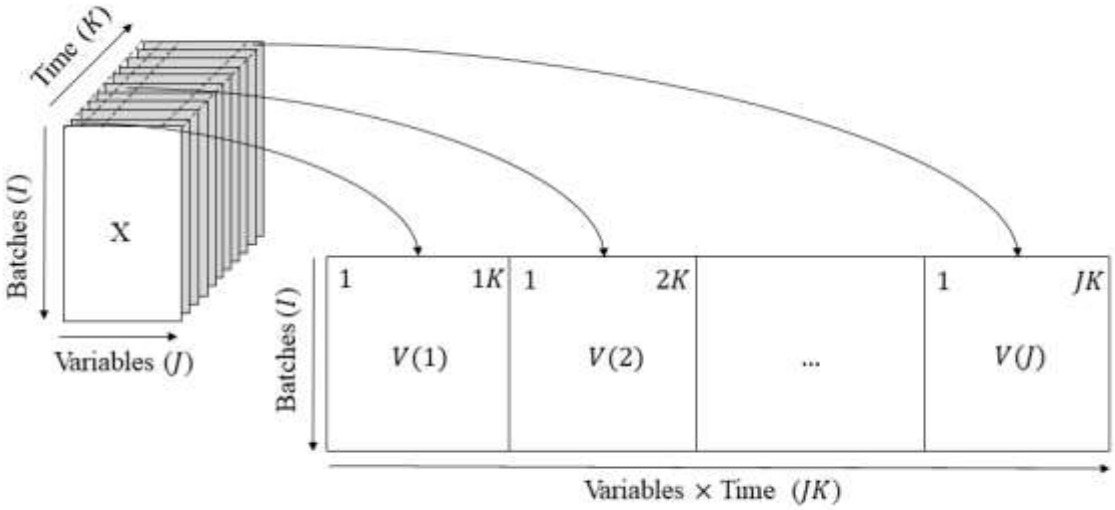


Figure 4. 1 Unfolding of the three-dimensional process data matrix

Prior to obtaining a PCA model each column of $\mathbf{X}_{\boldsymbol{B}}$ should be standardized to display zero mean and unit variance; let $\mathbf{Z}_{\boldsymbol{B}}$ denote the matrix of standardized variables. The purpose of standardization was to remove scale effects present when variables display means of different magnitude (KOSANOVICH; DAHL; PIOVOSO, 1996; ZHAOMIN; QINGCHAO; XUEFENG, 2014).

Analogously to the ordinary two-way PCA, MPCA rewrites the original datasets into a new JK–dimensional variable space, with coordinate directions known as principal components (PCs). Usually, most relevant process information is contained in a subset of R PCs and the PCA-model used to represent process variation has reduced dimension if compared to the original dataset (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; KHOSRAVI; MELÉNDEZ; COLOMER, 2009; KOSANOVICH; DAHL; PIOVOSO, 1996). Thus, the MPCA model rewrites the $\mathbf{Z}_{\boldsymbol{B}}$ matrix as the product of a score matrix $\mathbf{T}_{\boldsymbol{R}}(I \times R)$ and a transposed loading matrix $\mathbf{P}'_R(R \times JK)$, plus a residual matrix $\mathbf{E}(I \times JK)$. Whenever $R < JK$, dimension reduction is accomplished and the model in eqn. (1) is used to represent the original dataset without significant information loss (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; KHOSRAVI; MELÉNDEZ; COLOMER, 2009; WOLD, 1978; ZHAOMIN; QINGCHAO; XUEFENG, 2014).

$$\mathbf{Z}_{\boldsymbol{B}} = \mathbf{T}_R \times \mathbf{P}'_R + \mathbf{E} \qquad (1)$$

To project information contained in vector $\mathbf{x}_{new}$ obtained from a new batch into the model subspace in eqn. (1), the scores for the new batch are obtained as follows (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018):

$$\mathbf{t}_{new} = \mathbf{x}_{new} \times \mathbf{P}_R \qquad (2)$$

Once scores have been computed, a vector of residuals $\mathbf{e}_{new}$ (eqn. 3) is calculated as the difference between observed data and model estimates:

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \hat{\mathbf{x}}_{new} = \mathbf{x}_{new} - \mathbf{t}_{new} \times \mathbf{P}_R{}' \qquad (3)$$

where $\hat{\mathbf{x}}_{new}$ is an estimate based on the MPCA model (KOSANOVICH; DAHL; PIOVOSO, 1996).

The most popular criterion to decide on the number of retained PCs is the Kaiser-Guttman, in which components displaying eigenvalues larger than 1 are retained (JACKSON, 1993; RENCHER, 2002). When the mean of eigenvalues is larger than 1 (for example, when a covariance matrix is diagonalized), a variation of the criterion is adopted and all PCs for which $\lambda_r \geq \bar{\lambda}$ (λ_r is the eigenvalue associated with the r^{th} PC and $\bar{\lambda}$ is the mean value of the eigenvalues) are retained (REA; REA, 2016)

Scores in eqn. (2) are used to compute the Hotelling T^2 statistic (for the model space) and residuals in eqn. (3) are used to compute the Q statistic (for the residual space), as described next (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; KHOSRAVI; MELÉNDEZ; COLOMER, 2009; KOSANOVICH; DAHL; PIOVOSO, 1996).

(i) The Hotelling control chart statistic T^2_{new} is obtained projecting the scores $\mathbf{t}_{new}$ on the new set of orthogonal axes given by the R PCs retained in the reference data matrix $\mathbf{Z}_{\boldsymbol{B}}$, as follows:

$$T^2_{new} = \sum_{r=1}^{R} \frac{t_{new,r}{}^2}{\lambda_r} \qquad (4)$$

where $t_{new,r}$ is the score obtained from the r^{th} PC, and λ_r is its associated eigenvalue. Statistic T^2_{new} measures the fit of new observations to the reference model space and monitors the behavior of known sources of process variability (i.e. deviations in variables' time trajectories with respect to their reference trajectories).

(ii) The Q control chart statistic Q_{new} is used to detect unusual events that affect the correlation and autocorrelation structure captured by the reference PCA model. It is calculated as follows:

$$Q_{new} = \sum_{j=1}^{JK} (e_j^{new})^2 \qquad (5)$$

where e_j^{new} is the residual value associated with the jk^{th} unfolded variable in the new observation.

Process monitoring using PCA-based CCs is carried out in two phases. In Phase I, conforming batches selected from a historical database are analyzed such that process in-control (IC) behavior is characterized through a reference model. In Phase II, deviations in future batches' behavior with respect to the IC model are detected (CAPIZZI, 2015; DE OLIVEIRA; MARCONDES FILHO, 2018; FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; JONES-FARMER et al., 2014). Process monitoring may take place on-line (batch progress is monitored as it evolves), or off-line (batch behavior is verified at the end of the run). Both types of monitoring present advantages. On-line monitoring allows the implementation of corrective actions before the end of the run. For that, an out-of-control signal diagnosis must be available, which may not be possible in some types of process (e.g. rubber processing in which batches typically last 5 minutes or less). Off-line monitoring, on the other hand, allows characterizing final product quality several hours before lab test results become available, based solely on batch behavior. Independent of the control strategy, fault detection relates to the signaling of abnormal situations, fault isolation relates to the decomposition of the multivariate out-of-control signal and identification of out-of-control variables, and fault diagnosis relates to the identification of the mechanism responsible for the out-of-control signal (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; KHOSRAVI; MELÉNDEZ; COLOMER, 2009; KOURTI, 2003).

Potential abnormal events are acting on the process whenever the CCs statistics T^2 and Q produce points outside the charts' in-control region, bounded by their control limits (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; ZHAOMIN; QINGCHAO; XUEFENG, 2014). Upper and Lower Control Limits (UCL and LCL) for the Hotelling T^2 CC to be used in Phase I of process monitoring are given in eqn. (6), considering individual observations $(n = 1)$ and a significance level α (NOMIKOS; MACGREGOR, 1995b).

$$\mathrm{UCL}_{T^2,\mathrm{Phase\ I}} \leq \frac{(CF-1)^2}{(CF)} \times \frac{(R/(CF-R-1))\times F_{\alpha,R,(CF-R-1)}}{1+(R/(CF-R-1))\times F_{\alpha,R,(CF-R-1)}} \tag{6}$$

$$\mathrm{LCL}_{T^2,\mathrm{Phase\ I}} = 0$$

where CF is the number of conforming batches, R is the number of retained PCs, and $F_{\alpha,R,(CF-R-1)}$ is the tabled value of an F distribution with R and $(CF - R - 1)$ degrees of freedom and confidence limit α.

A common strategy in Phase I consists of iteratively revising control limits each time an out-of-control (*OOC*) batch is detected. Start iterations setting $CF = I$ (i.e. the total number of batches in $\mathbf{Z}_{\boldsymbol{B}}$) and eliminate batches yielding T^2 and/or Q values outside the charts' in-control region (eqns. 6 and 8). Let OOC denote a variable that counts the number of batches yielding out-of-control signals. Continue to iterate setting $CF = I - OOC$, and carry on the process until all batches yield CCs statistics values within control limits. Phase I results in (i) a reduced $\mathbf{Z}_{\boldsymbol{B}}$ dataset comprised only of IC batches which may be used to obtain an IC reference model through PCA, and (ii) estimates of unknown process parameters which may be used to calculate control limits for CCs in Phase II (CAPIZZI, 2015).

Phase II control limits for the T^2 CC considering $n = 1$ and a significance level α are given by (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018; LOWRY; MONTGOMERY, 1995):

$$\mathrm{UCL}_{T^2,\text{Phase II}} \leq \frac{R(CF^2-1)}{CF(CF-R)} \times F_{R,(CF-R),\alpha} \qquad (7)$$

$$\mathrm{LCL}_{T^2,\text{Phase II}} = 0$$

where R is the number of retained PCs, I is the number of batches, $CF = I - OOC$, and $F_{R,(CF-R),\alpha}$ is the tabled value for an F distribution with R and $(CF - R)$ degrees of freedom and confidence limit α. T^2_{new} values outside control limits potentially represent a situation in which at least one variable in $\mathbf{x}_{new}$ significantly deviates from its reference trajectory (KHOSRAVI; MELÉNDEZ; COLOMER, 2009).

Monitoring of a new batch using the Q chart is carried out using eqns. (2), (3) and (5). Control limits for the Q chart in Phase II are given by (FUENTES-GARCÍA; MACIÁ-FÉRNANDEZ; CAMACHO, 2018):

$$\mathrm{UCL}_Q \leq \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \qquad (8)$$

$$\mathrm{LCL}_Q = 0$$

where $\theta_1 = \sum_{i=R+1}^{J} \lambda_i$, $\theta_2 = \sum_{i=R+1}^{J} \lambda_i^2$ and $\theta_3 = \sum_{i=R+1}^{J} \lambda_i^3$, such that λ_i is the eigenvalue associated with the i^{th} PC, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, and z_α is the standard normal distribution value for a false alarm probability α. Q_{new} values outside control limits indicate that atypical events changed the correlation/auto-correlation structure represented in the reference model (DE OLIVEIRA; MARCONDES FILHO, 2018).

4.2.2 Variable selection based on accuracy classification of batches

Consider an $\mathbf{Z}_{\boldsymbol{B}}(I \times JK)$ input dataset and an $\mathbf{y}\,(I \times 1)$ output vector giving the outcomes of a binary categorical variable that discriminates between conforming and non-conforming batches. Batches are randomly split into training (I_{tr}) and testing (I_{ts}) sets, such that $I_{tr} + I_{ts} = I$. Batches in the training set are used to select the optimal subset of process variables using a PVS method that considers two criteria: classification accuracy and percentage of retained variables (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012). To obtain the relationship between JK process variables and the categorical output variable, PLS-DA is implemented. PLS-DA is recommended for cases in which $I < JK$, i.e. the number of observations (batches) is smaller than the number of JK unfolding variables (YAN; KUANG; YAO, 2017). In addition, a variable selection based on PLS-DA relies on variables that best discriminate IC and OOC batches, which is in agreement with the fault detection requirement of assigning samples in two classes, conforming and non-conforming (KUANG; YAN; YAO, 2015).

In PLS-DA each entry y_{ig} of $\mathbf{y}$ represents the g^{th} class of the i^{th} observation expressed with a binary code (1 or 0). In two-class categorization the number 1 usually indicates that the training observation belongs to the class of interest (conforming batches), while 0 indicates that the observation belongs to a different class (non-conforming batches). However, estimates of $\mathbf{y}$ obtained from a PLS-DA model are continuous, non-integer values, and a threshold should be defined for each class; e.g. whenever y_{ig} is greater than or equal to the threshold defined for the g^{th} class, sample i will be assigned to the g^{th} class; otherwise it will be assigned to the other class. The simplest approach for a two-class categorization is to use an arbitrary cut-off value such as 0.5 (BALLABIO; CONSONNI, 2013; PÉREZ; FERRÉ; BOQUÉ, 2009).

PLS-DA output parameters $\mathbf{w}, \mathbf{p}_a$ and E_{Ya}^2 are used to compute the importance index VII_j for process variable j [eqn. (9)]. The weight vector $\mathbf{w}$ that defines the linear combination of process variables is selected to maximize the covariance with $\mathbf{t}_{new}$. The fraction of the variance in $\mathbf{y}$ explained by the retained component a is given by E_{Ya}^2. For each component a, a vector of loadings $\mathbf{p}_a$ is obtained regressing the columns of $\mathbf{Z}_{\boldsymbol{B}}$ on $\mathbf{t}_{new}$ (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012). The number of PLS-DA components to be retained, A, could be defined through cross-validation and does not need to be equal to R's number of components retained by the PCA used for estimating the T^2_{new} statistic eqn (4).

$$VII_j = \sum_{a=1}^{A} \left(\frac{w_{ja}}{p_{ja} \times w_{ja}}\right)^2 E^2_{Ya} \qquad (9)$$

where $j = 1, \ldots, JK$.

VII_j results are organized in a descending order to create candidate sets (CS) of selected process variables. Classification accuracy (CA) of each CS [eqn. (10)] will be determined using the k-nearest Neighbor (kNN) classification technique (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012).

$$CA = \frac{I_{CC}}{Class_{Total}}, \qquad (10)$$

where I_{CC} is the number of correctly classified batches and $Class_{Total}$ is the total number of classifications. The k-nearest neighbor is a data mining technique which splits the I_{tr} subset into training ($I_{tr\prime}$) and testing portions ($I_{ts\prime}$), and classifies new observations based on the class labels of the k nearest neighbors (i.e. the ones with smallest Euclidean distance) in the variable space. When the class of each observation in I_{tr} is known beforehand, the value of k is the one that maximizes the classification accuracy in the testing portion $I_{ts\prime}$ (DUDA; HART; STORK, 2001; WU et al., 2008).

An iterative procedure involving VII and kNN is run. First, the classification accuracy of $I_{ts\prime}$ is computed using all JK variables. In the following iterations, using backward elimination each succeeding CS is obtained by removing the variable with the lowest VII_j from the set, until there is a single remaining variable (RV); see eqn. (11). At each step, the classification accuracy is computed.

$$RV = JK - j, \qquad (11)$$

where $j = 1,2, \ldots, JK - 1$

To handle the large number of CS when the number of process variables is very large, a Pareto Optimality (PO) analysis may be used. The idea is to determine a group of sets, called the non-dominated candidate sets (CS_{nd}), such that no other sets in the search space are superior to them. Then, the CS_{nd} with smallest Euclidean distance from a hypothetical ideal set, which minimizes the number of retained variables and maximizes classification accuracy, is chosen and named Pareto Variable Selection (PVS) set. Finally, the performance of the PVS set is verified in the I_{ts} subset (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012; TABOADA; COIT, 2007; ZITZLER; THIELE, 1999).

4.3 PVS-MPCA FAULT DETECTION METHOD

The PVS-MPCA monitoring method consists of combining a wrapper variable selection aimed at reducing the original dataset to a subset of relevant variables for batch classification, previous to the construction of the T^2 and Q CCs for fault detection in high dimensional datasets. A limitation of MPCA-based CCs applied to batch processes is that batches must have the same duration (MARTIN; MORRIS; KIPARISSIDES, 1999), which is rarely the case in industrial applications. Therefore, batch data must be pre-processed following one of the time warping strategies presented by Peres et al. (2018), that promote alignment and synchronization of batches.

Figure 4.2 presents an overview of the PVS-MPCA monitoring method. Method steps are explained next.

(1) Pareto Variable Selection

Step 1: Split the dataset into I_{tr} (90%) and I_{ts} (10%) portions.

Step 2: Run a PLS-DA on I_{tr} and obtain parameters w_{ja}, p_{ja}, and E_{Ya}^2. The total number of retained PCs (A) is determined using a 10-fold cross-validation procedure (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

Step 3: Determine the VII_j of each process variable in the dataset and organize them in descending order.

Step 4: Split I_{tr} into training ($I_{tr'} = 80\%$) and testing ($I_{ts'} = 20\%$) portions. Train the kNN algorithm using $I_{tr'}$ and compute the classification accuracy on $I_{ts'}$. Iterate this training/testing procedure twenty times to obtain the average accuracy classification (ACA) at each iteration. For the first candidate set CS, use $RV = JK$; in the next iteration, remove the variable with the smallest VII_j value and obtain the next CS. When the stopping criterion is achieved, plot $\%RV$ *versus* ACA. The number of neighbors k is determined using a 10-fold cross-validation based on classification performance metrics; namely accuracy, specificity and sensitivity. Sensitivity is determined dividing the number of true conforming batches by the total number of batches classified as conforming; specificity is determined dividing the number of true non-conforming batches by the total number of batches classified as non-conforming (ALPAYDIN, 2010). In case of a tie, choose the smallest k.

Step 5: Define a Pareto frontier to reduce the number of candidate sets and select the CS closest to the hypothetical ideal set, which minimizes the RV and maximizes the

ACA; in this paper, we use $RV = 0.00035$ (only 1 of 2,864 variables retained) and ACA = 1. The selected CS is the PVS set.

Step 6: Compare the ACA in the I_{ts} portion using only the PVS subset of selected variables with the ACA obtained in I_{ts} using all variables.

(2) Phase I – Build a reference distribution using the MPCA monitoring method

Step 7: Consider an input dataset $\mathbf{Z}_{\boldsymbol{B}}(I_{CFI} \times PVS)$ where I_{CFI} are batches classified as conforming by the industry. Run a PCA and compute T^2 and Q statistics for each batch, as well the CCs' control limits. Using the iterative procedure for Phase I described in section 4.2.1, obtain the reference set of conforming batches, develop the PCA monitoring model for the reference distribution, and determine $I_{CFI,NCF}$; i.e. the group of batches deemed conforming by industry classification but flagged as non-conforming in the CCs. Batches in $I_{CFI,NCF}$ are removed from the reference distribution, and used to compute FAR (eqn. 13).

(3) Phase II – Offline MPCA Fault Detection

Step 8: Compute CCs' control limits for Phase II. Calculate T^2_{new} and Q_{new} statistics for each batch in I_{NCFI}, which is comprised of batches deemed non-conforming by the industry. Plot batches on the CCs and run fault detection. Batches beyond control limits in any of the charts are flagged as non-conforming, belonging to set I_{NCF}, while batches within control limits are considered conforming, belonging to set I_{CF2}. To evaluate the monitoring performance of the CCs the cumulative error rate (GHOSH; RAMTEKE; SRINIVASAN, 2014) is determined as follows:

$$Cumulative\ Error\ Rate = FAR + MDR \tag{12}$$

$$FAR = 100 * \frac{I_{CFI,NCF}}{CFI_{Total}} \tag{13}$$

$$MDR = 100 * \frac{I_{NCFI,CF2}}{NCFI_{Total}} \tag{14}$$

where CFI_{Total} is the total number of conforming batches (according industry classification), $I_{NCFI,CF2}$ is the total number of non-conforming batches (according industry classification) flagged as conforming in Phase II, and $NCFI_{Total}$ is the total number of non-conforming batches (according industry classification).

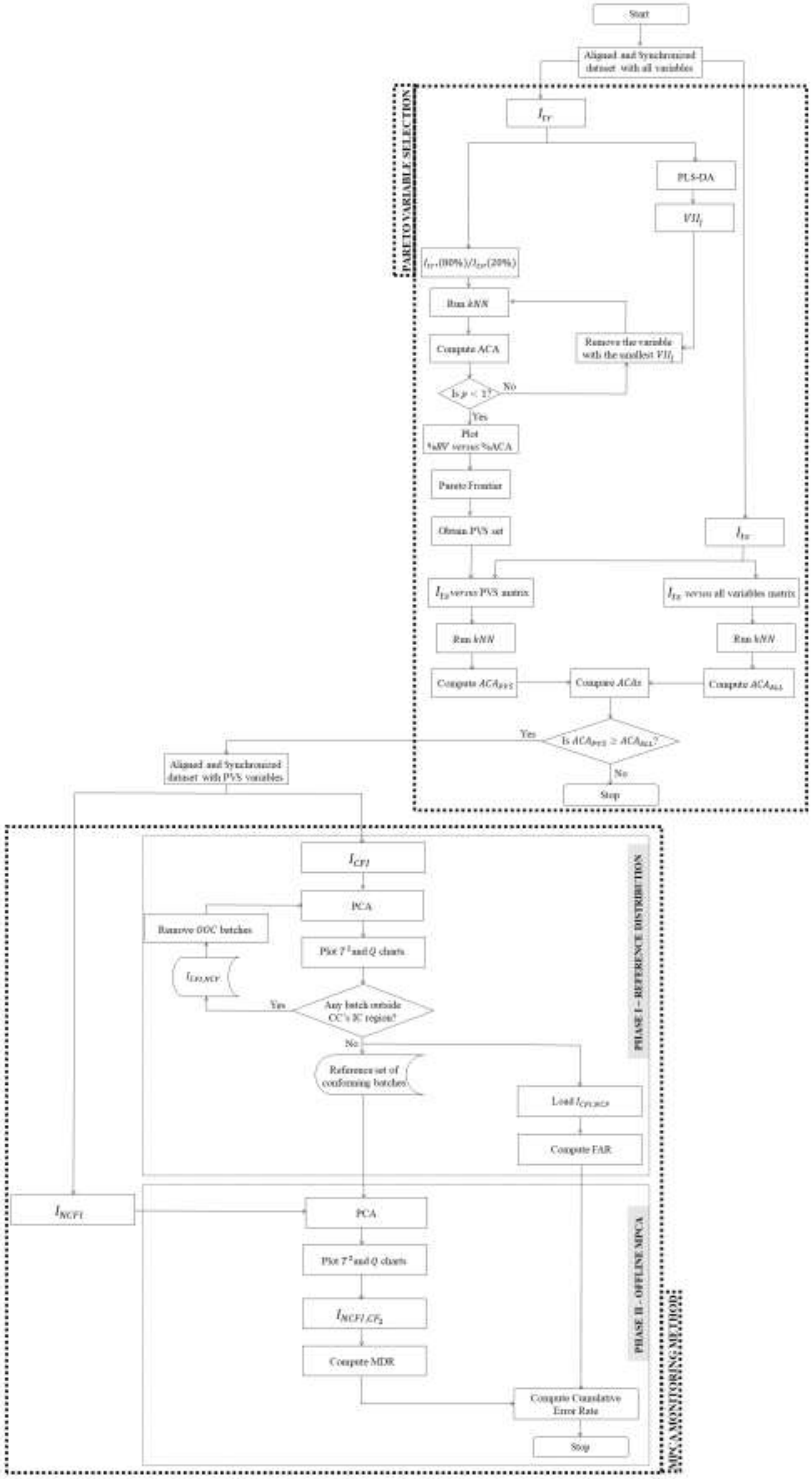

Start
Aligned and Synchronized dataset with all variables
PARETO VARIABLE SELECTION
I_{TT}
PLS-DA
VII_j
I_{TT}(80%)/I_{TV}(20%)
Run kNN
Compute ACA
Remove the variable with the smallest VII_j
Is $p < 2$?
No
Yes
Plot %RV versus %ACA
Pareto Frontier
Obtain PVS set
I_{TS}
I_{TS} versus PVS matrix
I_{TS} versus all variables matrix
Run kNN
Run kNN
Compute ACA_{PVS}
Compare ACAs
Compute ACA_{ALL}
Yes
Is $ACA_{PVS} \geq ACA_{ALL}$?
No
Stop
Aligned and Synchronized dataset with PVS variables
I_{CF1}
PCA
Remove OOC batches
Plot T^2 and Q charts
$I_{CF1,NCF}$
Yes
Any batch outside CC's IC region?
No
Reference set of conforming batches
Load $I_{CF1,NCF}$
Compute FAR
PHASE I – REFERENCE DISTRIBUTION
I_{NCF1}
PCA
Plot T^2 and Q charts
I_{NCF1,CF_2}
Compute MDR
Compute Cumulative Error Rate
Stop
PHASE II - OFFLINE MPCA
MPCA MONITORING METHOD

Figure 4. 2 PVS-MPCA method – implementation steps

4.4 CASE STUDY AND DISCUSSION

4.4.1 Process description

Process data obtained from batches of chocolate submitted to a conching operation were made available by a large chocolate manufacturing plant. The dataset is comprised of 69 batches of milk chocolate processed between April 2014 and January 2015. Process variables were sampled at 1-minute intervals during the production of each batch. Process data from the same milk chocolate recipe were obtained from 22 consecutive batches processed in a Frisse conche type ELK (Bühler AG, Uzwill, Switzerland), named conche A, and 47 consecutive batches processed in another Frisse conche type ELK, named conche B. Both conches have the same operational design and operate in a similar way.

The cycle in the conche consists of four phases, namely: feeding, dry conching, plastic conching, and liquefaction (BÜHLER, 2010). Refined chocolate is loaded into the conche and temperature is raised up to $45°C$ during the feeding phase. After all refined chocolate is charged dry conching takes place, in which the chocolate mass is mixed under high temperature $(70 - 80°C)$ to promote the evaporation of water and undesirable volatiles (BOLENZ; KUTSCHKE; LIPP, 2008; BÜHLER, 2010; PRAWIRA; BARRINGER, 2009). Then the plastic phase begins with the addition of fat and/or soybean lecithin emulsifier to the chocolate mass (BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; GLICERINA et al., 2015), allowing the current dependent drive to change from clockwise to anti-clockwise rotation. In order to start the liquefaction phase, chocolate mass temperature is reduced to $45°C$ and vanilla flavor and emulsifiers (soybean lecithin and/or polyglycerol polyricinoleate) are added, while intense shearing is applied. Upon conclusion of this phase, a fluid chocolate mass should be available with adequate rheological properties (AFOAKWA et al., 2008; BOLENZ; THIESSENHUSEN; SCHÄPE, 2003; BÜHLER, 2010; GLICERINA et al., 2013). To verify that, a sample of the mass is collected and its rheological parameters are assessed. If results are not in accordance with industrial standards, a new addition of emulsifiers and/or vegetable fat is usually required to correct the chocolate mass rheology.

The expected batch duration is 495 minutes, divided as follows: 60 minutes of feeding, 150 minutes of dry conching, 150 minutes of plastic conching, 120 minutes of liquefaction and 15 minutes for discharging (BÜHLER, 2010). However, the real duration is variable and batches may take up to 1,170 minutes of processing for completion. Batch duration above the expected usually occurs when rheological specifications are not achieved, requiring corrections in the chocolate mass.

To ensure that all batches have the same total duration the dataset was pre-treated as described in section 4.3. After alignment and synchronization, a dataset of 62 batches (I) in which 4 process variables (J) were measured 716 times (K) was available. The list of J process variables is shown in Table 4.1. The classification of each batch was based on off-line viscosity results; laboratory measurements of this single quality output were made after batch completion, and results were available 15 minutes after sample collection. Using the plant threshold viscosity value, batches were classified as conforming (56%) or non-conforming (44%).

Table 4. 1 Process variables measured during batch progression

Variable name	**Variable description**
Metering system n°6	Frequency of soybean lecithin dosage
Speed	Rotation speed of conche shovels
Drive current	Current of the conche's main motor
Chocolate temperature	Temperature of chocolate mass during the conching

4.4.2 Variable selection

To find the best subset of process variables for classification of production batches in conforming and non-conforming, the PVS method was applied to $\mathbf{Z}_{\boldsymbol{B}}$ (62 batches × 2,864 variables) and $\mathbf{y}$ (62 batches × 1 categorical quality parameter). Applying a 10-fold cross-validation, 8 components were retained to build the PLS-DA regression model. The dataset was randomly split in 90% training (I_{tr}) and 10% testing (I_{ts}) portions, and PLS-DA regression was applied in the training portion. To avoid basing VII_j results on only one partition of the dataset, 12,000 PLS-DA models were run to obtain convergence in importance indices; the final ordering of process variables was named $VII_{converged}$. The PLS-DA partition that best resembled the ordering of variables in $VII_{converged}$ was chosen; 95.92% of the variation in $\mathbf{y}$ was captured by the retained PCs.

To define the value of parameter k in the kNN classification technique a 10-fold cross-validation was applied; $k = 3$ yielded the highest average accuracy, sensitivity, and specificity across all tested values. The PVS method was applied to minimize the number of RV (retained variables) and maximize the ACA.

Results obtained during the iterative process between CS and kNN classification technique are displayed in Figure 4.3.

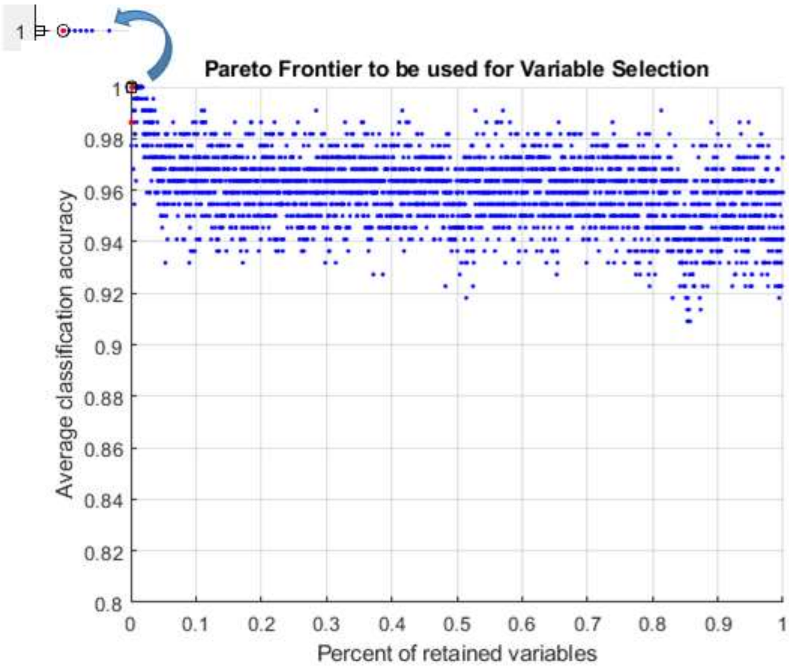


Figure 4. 3 Average accuracy classification for different candidate sets of variables

The Pareto frontier is signalized as red points, and the hollow circle represents the CS_{nd} selected by the PVS method as the closest solution to the hypothetical ideal point (hollow square in detail). The PVS solution is composed of 5 variables yielding 100% ACA. Considering that with the complete set of 2,864 variables the ACA was 95.91%, retaining only 0.17% of them greatly improved the power of classification of batches in conforming and non-conforming, simplifying the CC analyses to follow.

Retained variables and their time periods of measurement are shown in Table 4.2. Variable ‘Metering system n°6’ was not retained in any time period in which it was measured during batch progression. Remaining three variables were retained in five time periods; note that measurements of ‘Chocolate temperature’ at the beginning and end of batches were retained.

To finish implementing the method, the I_{ts} portion was classified using only the variables in the PVS set, yielding 100% of ACA. This result was the same obtained running the classification on I_{ts} using all JK original variables. This is an important result, since it is highly desirable not to lose the classification ability after variable selection, since correct classification is key for the construction of a reference model and the monitoring of future batches.

Table 4. 2 List of retained process variables and their time of measurement during batch progression

Variable name	**Time period**	**Step of conching process**
Drive Current	68	Feeding
Speed	203	Dry conching
	204	Dry conching
Chocolate temperature	1	Feeding
	716	Liquefaction

4.4.3 Offline fault detection

In this section, results of fault detection based on traditional MPCA on the dataset including all variables were compared to those obtained using the PVS-MPCA method on selected variables.

According to the manufacturer specifications 35 batches were considered conforming. Thus, the performance of MPCA in Phase I was evaluated based on the dataset $\mathbf{Z}_{\boldsymbol{B}}$ (35 batches $\times$ $2{,}864$ variables), following an iterative procedure in which batches with chart points falling beyond control limits were removed one at a time. After several iterations the reference distribution was obtained using only 8 batches identified as IC by T^2 and Q CCs (Figure 4.4). Two PCs were retained considering a modified Kaiser-Guttman criterion (REA; REA, 2016), explaining 74.75% of the variance in the process. Data were normalized, as discussed in section 4.2.1, prior to analysis.

In Phase II of traditional MPCA monitoring all batches classified as non-conforming by the industry were rescaled using the sample mean and variance obtained from the IC data. They were then projected onto the loading matrix to obtain the corresponding PC scores, and monitoring statistics T^2 and Q were computed. Figure 4.5

shows that during Phase II all batches were considered conforming in the T^2 chart; however, all batches were tagged as non-conforming in the Q residuals CC.

Based on T^2 and Q CC results 27 conforming batches were wrongly discarded as non-conforming in Phase I, totalizing 77.14% of FAR. All non-conforming batches were correctly signalized in Phase II representing a null MDR. The cumulative error rate computed for MPCA method was therefore 77.14%.

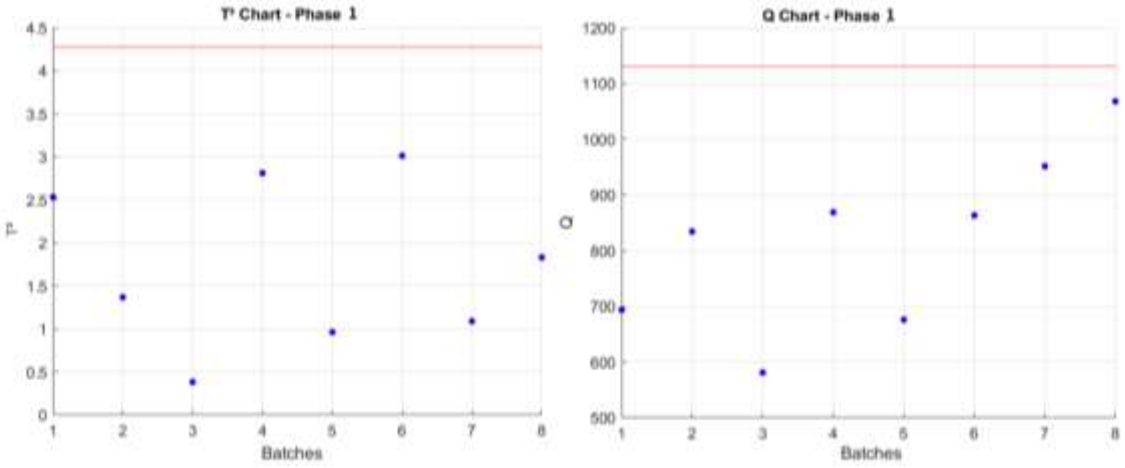


Figure 4. 4 Final iteration of Phase I of the MPCA method

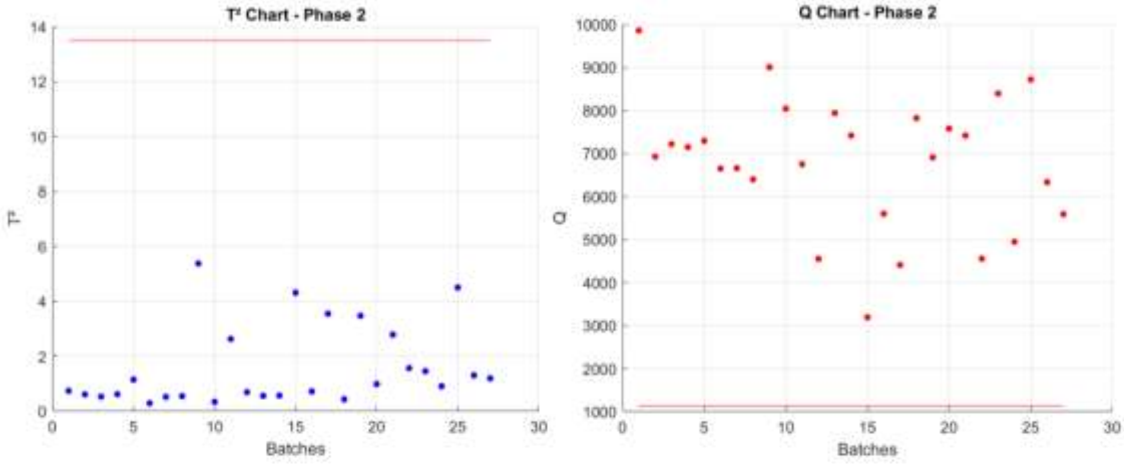


Figure 4. 5 Monitoring of new batches in Phase II using the MPCA method

We now present the results of applying the proposed method to the same dataset analyzed above. Phase I of PVS-MPCA was evaluated based on the dataset of $\overline{\mathbf{X}}_{\boldsymbol{B}}$ (35 batches $\times$ 5 variables), and the reference model was built using 31 batches identified as IC by T^2 and Q CCs (Figure 4.6), after some iterations. Two principal components were retained using the Kaiser-Guttman criterion; they accounted for 60.83% of the total variance in the dataset. All 4 batches were eliminated based on T^2 CC results, indicating deviations of process variables with respect to their reference trajectories.

In Phase II 3 non-conforming batches were wrongly identified as conforming in the T^2 CC, but correctly signalized as non-conforming in the Q CC. All other new batches were correctly signalized as non-conforming in both charts (Figure 4.7).

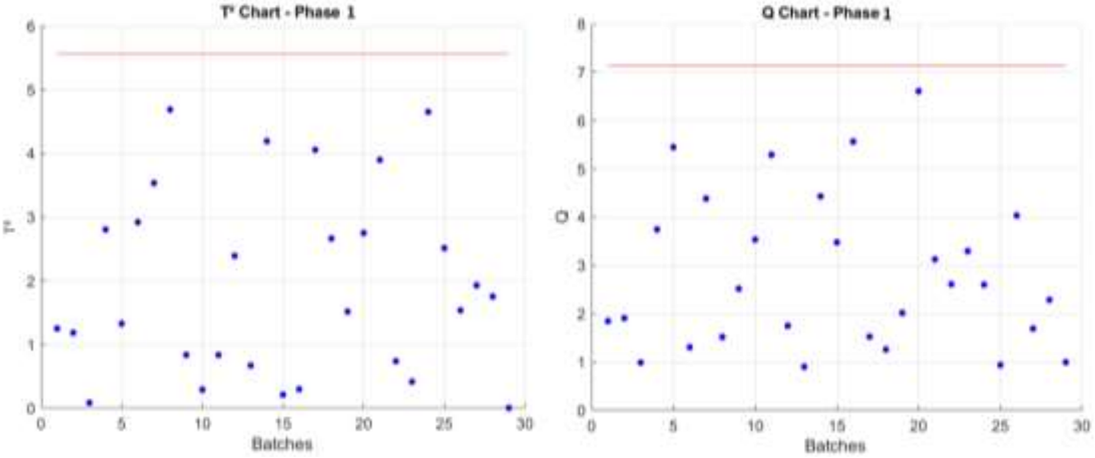


Figure 4. 6 Final iteration of Phase I using the PVS-MPCA method

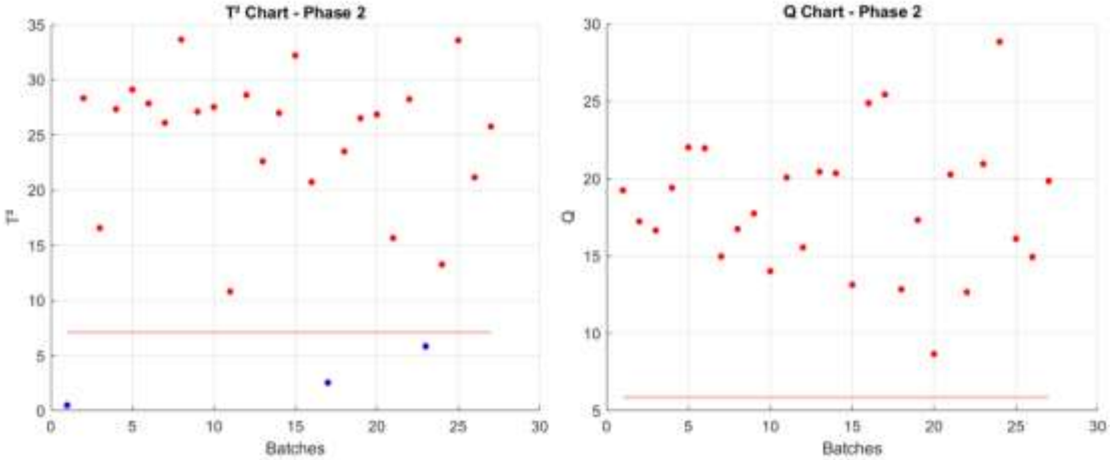


Figure 4. 7 Monitoring of new batches in Phase II using the PVS-MPCA method.

The 4 conforming batches wrongly detected as non-conforming in Phase I represented 11.43% of FAR. Based on T^2 and Q CCs results, all non-conforming batches were correctly signalized during Phase II, representing a null MDR. The cumulative error rate computed for PVS-MPCA method was therefore 11.43%.

4.4.4 Discussion

PCA promotes data dimensionality reduction by finding PCs that are linear combinations of the original variables in the dataset and keeping only a portion of them for further use. Such simple strategy has been challenged by modern applications where the massive number of process variables compromises the interpretation of retained PCs, particularly in the context of fault detection, isolation, and diagnosis.

The application of a standard MPCA to the high dimensional dataset obtained from the chocolate conching batch process highlighted the method's limitations. First, eigenvalues obtained in the analysis were very large in scale. When eigenvalues are obtained through diagonalization of a correlation matrix (which is the equivalent of obtaining eigenvalues from the covariance matrix of a normalized dataset) their mean is equal to 1, which is the basis of the Kaiser-Guttman criterion to decide the number of retained PCs (JACKSON, 1993; RENCHER, 2002). In the chocolate dataset, even though MPCA was performed on the correlation matrix the mean of the eigenvalues was 82.7 and all eigenvalues (except for one) were greater than 1, with the first seven being larger than 100. That could be attributed to the poor estimates resulting from using the sample correlation matrix when the number of variables is much larger than the number of observations (AMINI, 2011; WANG; FAN, 2017). That corroborates the recommendation in Stevens (2009) *apud* Gajjar, Kulahci, and Palazoglu (2018) of applying the Kaiser-Guttman criterion only in datasets with less than 30 variables and at least 250 observations.

To verify the impact of obtaining standard MPCA-based CCs using high dimensional datasets, a variation of the Kaiser-Guttman criterion (REA; REA, 2016) was applied. The high FAR observed in Phase I for T^2 and Q CCs, probably due to the use of a wrong IC model, proved MPCA to be unreliable for the monitoring of high dimensional processes. Capizzi (2015) and Jones-Farmer et al. (2014) recommended process characterization using a large number of observations (batches) and data sampling covering long periods, capturing both long- and short-term process characteristics and achieving reasonably accurate parameter estimates to be use in Phase II of process monitoring. In addition, MacGregor and Kourti (1995) recommended to use 50 or more conforming batches to obtain a representative sample and correctly estimate confidence limits for the normal operating region. However, in our case study it would be infeasible to solve the MPCA drawbacks reported above by increasing the number of batches, since the number of variables is over 40 times larger than the number of batches.

In Phase II all new batches were labeled conforming by the T^2 chart; and non-conformities appeared in the Q CC. That indicates low variability in the means within PCs, revealing the inability of the MPCA model to describe significant process variation. The Q statistic displayed large values, which indicates that it was significantly affected by process noise. According to Nomikos and MacGregor (1995b), in offline monitoring the Q statistic in a certain time period accumulates values from previous periods, and large squared residuals tend to be levelled out by small squared residuals over time. In our case, the large Q statistic values indicate an important lack of fit of non-conforming data to the MPCA model.

There is an increasing number of methods proposed to solve problems attributed to running MPCA on high dimensional datasets; they integrate variable selection to MSPC methods. However, the majority of them are proposed for continuous processes (PERES; FOGLIATTO, 2018), and only a few address batch processes. Some examples of integration of variable selection and batch process monitoring methods are described next.

Two early methods by Zarzo and Ferrer (2004) aimed to use variable selection based on technical knowledge to find variables that most contributed to model prediction performance in batch processes for later use in fault diagnosis. In the first, MPLS regression was applied to a batch dataset and weights were analyzed to distinguish groups that were related to process deviations. Technical knowledge was used to find an explanation for the observed correlation; whenever no reasonable explanation was found, variables were removed from the dataset. In the second method named Block-wise PCR, the unfolded matrix of batch process data was subdivided into blocks of variables' trajectories and a PCA was run in each block. A simple linear regression was conducted for every latent variable of a PC score matrix, considering one response variable. Based on the squared linear correlation coefficient and the *p*-value, variables significantly correlated with the response were selected. Then, expert knowledge was used to promote a second round of variable selection searching for latent variables displaying a change in trend in the CUSUM CC happening in parallel with changes in the response.

Chu, Qin, and Han (2004) performed a variable selection using an entropy measure and the sequential forward floating selection algorithm to determine variables that minimize the total entropy. These variables were used to compose a hyperspace in which different data clusters were identifiable. Next, Support Vector Machine (SVM) classifiers were constructed to define decision boundaries between normal and faulty data groups (and between different normal modes) without relying on the normality assumption. Such information was used as a criterion to conduct a hierarchical fault detection and operation mode identification.

Recently the improvement of fault isolation in batch processes was discussed by Yan, Kuang, and Yao (2017), who proposed the use of sparse Partial Least Squares (sPLS) to handle with (*i*) the asymptotic inconsistency of the PLS estimator in datasets with very large J and small I, (*ii*) autocorrelations and cross-correlations in batch process data, and (*iii*) identification of process critical variables when abnormalities have already been detected. sPLS builds a discriminant analysis model for normal and faulty operation batch data, achieving regression modeling and variable selection simultaneously; the order in which variables enter the sPLS model reflects their importance and the most critical time interval in a previously detected abnormal batch.

Zarzo and Ferrer (2004) based their propositions on the use of technical knowledge to select variables that most contributed to model prediction capacity. Their approach differs from sPLS (YAN; KUANG; YAO, 2017) and our proposed PVS-MPCA method that applied quantitative methods to select variables based on their ability to classify batches in conforming and non-conforming.

The SVM classification technique applied in Chu, Qin, and Han (2004)'s method was evaluated in the selection of the most important variables for product classification, resulting in higher number of retained variables and lower classification accuracy when compared to the kNN classification technique used in the PVS-MPCA method. SVM integrated to entropy-based variable selection applied for fault detection (CHU; QIN; HAN, 2004) and the sPLS-based fault isolation method (YAN; KUANG; YAO, 2017) are methods that do not use MPCA-based CCs for fault detection and isolation. In opposition, the PVS-MPCA method proposes an innovative way to extend the use of traditional T^2 and Q control charts for fault detection in high dimensional batch datasets.

When applying PVS-MPCA to the chocolate dataset 5 out of 2,864 unfolded variables were retained in the PVS set. They represent transition points in the original (unfolded) variables' trajectories. Aligned and synchronized batches were time-warped to last 716 minutes; of those, 133 minutes corresponded to the feeding operation, 141 to dry conching, 222 to plastic conching, and 220 to liquefaction. The variable ‘Drive current’ was selected at the time period when the current starts to increase during the feeding phase ($t = 68$). Variable ‘Speed’ was selected in time periods 203 and 204 when speed is reduced and then set to the maximum allowable speed until the chocolate mass becomes more fluid, requiring less power and speed to be mixed. Finally, variable ‘Chocolate temperature’ was selected at the beginning and at the end of the batch, where variations in the temperature profile are more evident. Warped trajectories of variables ‘Drive current’, ‘Speed’ and ‘Chocolate temperature’ are shown in Figure 4.8. Variable ‘Metering system n°6’ was not selected at any time period, suggesting that the instants in which the addition of soybean lecithin takes place are not determinant of the final classification of batches.

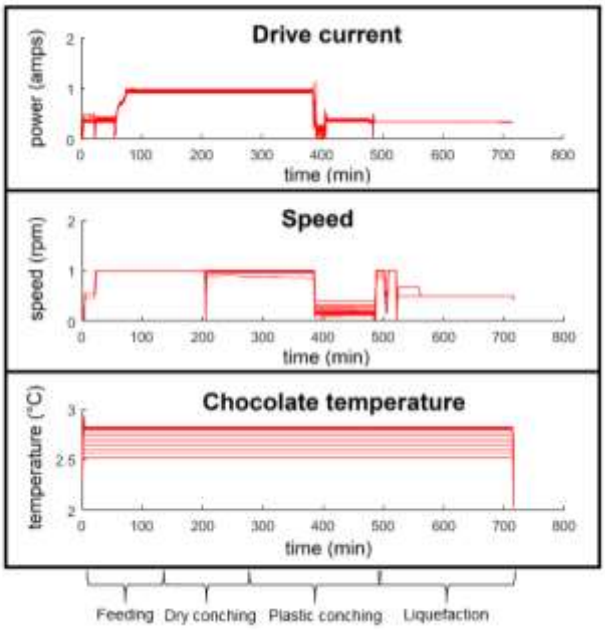


Figure 4. 8 Aligned and synchronized trajectories of original (unfolded) variables

The PVS-MPCA method for high dimensional datasets yielded a reduction of 85.18% in FAR (from 77.14% to 11.43%) when compared to traditional MPCA monitoring applied to the chocolate dataset. That greatly improves fault detection. Four conforming batches were wrongly signalized as non-conforming in Phase I of PVS-MPCA. In practice, that is not viewed as a significant problem since to adjust the rheology of an out-of-specification batch an additional viscosity analysis will be required; any misclassification of conforming batches would be detected at this point.

The 4 conforming batches eliminated in Phase I of PVS-MPCA were not produced in sequence. Two were produced in conche A, and two in conche B during the last period of data collection, from November to January. The conches clearly do not have an effect on accuracy of batch classification; however, outside temperature may influence the results. November, December and January are summer months in the plant's location, when the temperature may reach 40°C and affect the measurement quality of sensors.

In Phase II, Q values were much smaller in PVS-MPCA (Figure 4.7) than in MPCA monitoring (Figure 4.5) which indicates a better performance of MPCA model. The fact that MDR is 0 is of great practical importance since the misclassification of non-conforming batches leads to severe technical problems (e.g., imperfect covering of bonbons, poor spreading of chocolate mass into bar molds, and clogging of pipes and moving parts of the moulding equipment).

4.5 CONCLUSION

In this paper we present a new fault detection method in which a wrapper variable selection step is applied prior to the construction of MPCA-based T^2 and Q control charts. The performance of our proposed method was verified using real data from a chocolate conching batch process. The iterative procedure based on the variable importance index VII obtained using PLS-DA regression output parameters and the kNN classification technique retained only the subset of most discriminating variables, improving the accuracy of batch classification from 95.91% with 2,864 unfolded variables to 100% with only 5 variables. Control charts' performance using only variables in the PVS set reduced the false alarm rate by 85.18% when compared to CCs built using all variables.

PVS-MPCA clearly overcome the limitations of MPCA-based CCs when applied to high dimensional datasets. Moreover, it potentially improves fault isolation since variable selection greatly reduces the number of variables to be further investigated in the contribution plots. That will be the subject of future research on the method.

4.6 REFERENCES

AFOAKWA, E. O.; PATERSON, A.; FOWLER M.; VIEIRA J. Relationship between rheological , textural and melting properties of dark chocolate as influenced by particle size distribution and composition. **European Food Research Technology**, v. 227, p. 1215–1223, 2008.

ALPAYDIN, E. **Introduction to machine learning.** 2ed. Cambridge: The MIT Press, 2010. 537p.

AMINI, A. A. **High-dimensional principal component analysis**. 2011. 163p. Thesis (Doctorate) - Electrical Engineering and Computer Sciences, University of California, Berkley.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research,** v. 218, n. 1, p. 97–105, 2012.

BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. **Analytical Methods**, v. 5, p. 3790–3798, 2013.

BISGAARD, S. The future of quality technology: From a manufacturing to a knowledge economy & from defects to innovations. **Quality Engineering**, v. 24, n. 1, p. 30–36, 2012.

BOLENZ, S.; KUTSCHKE, E.; LIPP, E. Using extra dry milk ingredients for accelerated conching of milk chocolate. **European Food Research Technology**, v. 227, p. 1677–1685, 2008.

BOLENZ, S.; THIESSENHUSEN, T.; SCHÄPE, R. Fast conching for milk chocolate. **European Food Research Technology**, v. 218, p. 62–67, 2003.

BÜHLER, A. Operating instructions ELK / DÜC conches. Uzwill, Switzerland: [s.n.].

CAPIZZI, G. Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II. **Quality Engineering**, v. 27, p. 44-67, 2015.

CHU, Y.-H.; QIN, S. J.; HAN, C. Fault Detection and Operation Mode Identification Based on Pattern Classification with Variable Selection. I**ndustrial & Engineering Chemistry Research**, v. 43, p. 1701–1710, 2004.

DE OLIVEIRA, L. P. L.; MARCONDES FILHO, D. Monitoring batch processes with an incomplete set of variables. **International Journal of Advanced Manufacturing Technology**, v. 94, p. 2515–2523, 2018.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2ed. New York: John Wiley, 2001. 637p.

EMPARÁN, M.; SIMPSON, R.; ALMONACID, S.; TEIXEIRA, A.; URTUBIA, A. Early recognition of problematic wine fermentations through multivariate data analyses. **Food Control**, v. 27, n. 1, p. 248–253, 2012.

FUENTES-GARCÍA, M.; MACIÁ-FÉRNANDEZ, G.; CAMACHO, J. Evaluation of diagnosis methods in PCA-based Multivariate Statistical Process Control.

Chemometrics and Intelligent Laboratory Systems, v. 172, p. 194–210, 2018.

GAJJAR, S.; KULAHCI, M.; PALAZOGLU, A. Real-time fault detection and diagnosis using sparse principal component analysis. **Journal of Process Control**, v. 67, p. 112-128, 2018.

GHOSH, K.; RAMTEKE, M.; SRINIVASAN, R. Optimal variable selection for effective statistical process monitoring. **Computers and Chemical Engineering**, v. 60, p. 260–276, 2014.

GLICERINA, V.; BALESTRA, F.; DALLA ROSA M.; ROMANI, S. Rheological , textural and calorimetric modifications of dark chocolate during process. **Journal of Food Engineering**, v. 119, p. 173–179, 2013.

GLICERINA, V.; BALESTRA, F.; DALLA ROSA M.; ROMANI, S. Effect of manufacturing process on the microstructural and rheological properties of milk chocolate. **Journal of Food Engineering**, v. 145, p. 45–50, 2015.

JACKSON, D. A. Stopping rules in principal components analysis : A comparison of heuristical and statistical approaches. **Ecology**, v. 74, n. 8, p. 2204–2214, 1993.

JIANG, Q.; HUANG, B. Distributed monitoring for large-scale processes based on multivariate statistical analysis and Bayesian method. **Journal of Process Control**, v. 46, p. 75–83, 2016.

JOHNSTONE, I. M.; LU, A. Y. On consistency and sparsity for principal components analysis in high dimensions. **Journal of the American Statistical Association**, v. 104, n. 486, p. 682–693, 2009.

JONES-FARMER, L. A.; WOODALL, W. H.; STEINER, S. H.; CHAMP, C. W. An overview of phase I analysis for process improvement and monitoring. **Journal of Quality Technology,** v. 46, n. 3, p. 265–280, 2014.

KHOSRAVI, A.; MELÉNDEZ, J.; COLOMER, J. Classification of sags gathered in distribution substations based on multiway principal component analysis. **Electric Power Systems Research**, v. 79, n. 1, p. 144–151, 2009.

KOSANOVICH, K. A.; DAHL, K. S.; PIOVOSO, M. J. Improved process understanding using multiway principal component analysis. **Industrial & Engineering Chemistry Research**, v. 35, n. 1, p. 138–146, 1996.

KOURTI, T. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. **Annual Reviews in Control**, v. 27, p. 131–139, 2003.

KUANG, T. H.; YAN, Z.; YAO, Y. Multivariate fault isolation via variable selection in discriminant analysis. **Journal of Process Control**, v. 35, p. 30–40, 2015.

LEE, Y. K.; LEE, E. R.; PARK, B. U. Principal component analysis in very high-dimensional spaces. **Statistica Sinica**, v. 22, p. 933–956, 2012.

LOWRY, C. A.; MONTGOMERY, D. C. A review of multivariate control charts. **IIE Transactions**, v. 27, n. 6, p. 800–810, 1995.

LUO, L.; BAO, S.; MAO, J.; TANG, D. Fault detection and diagnosis based on sparse PCA and two-level contribution plots. **Industrial and Engineering Chemistry Research**, v. 56, n. 1, p. 225–240, 2017.

MACGREGOR, J. F.; KOURTI, T. Statistical process control of multivariate processes. **Control Engineering Practice**, v. 3, n. 3, p. 403-414, 1995.

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, v. 23, p. 35–44, 1999.

NOMIKOS, P.; MACGREGOR, J. F. Monitoring batch processes using multiway principal component analysis. **AlChE Journal**, v. 40, n. 8, p. 1361–1375, 1994.

NOMIKOS, P.; MACGREGOR, J. F. Multi-way partial least squares in monitoring batch processes. **Chemometrics and Intelligent Laboratory Systems**, v. 30, p. 97–108, 1995a.

NOMIKOS, P.; MACGREGOR, J. F. Multivariate statistical process control charts for monitoring batch processes. **Technometrics**, v. 37, n. 1, p. 41-59, 1995b.

PERES, F. A. P.; FOGLIATTO, F. S. Variable selection methods in multivariate statistical process control: A systematic literature review. **Computers and Industrial Engineering**, v. 115, p. 603–619, 2018.

PERES, F. A. P.; PERES, T. N.; FOGLIATTO, F. S.; ANZANELLO, M. J. Strategies for synchronizing chocolate conching batch process using dynamic time warping. Unpublished results, p. 1–21, 2018.

PÉREZ, N. F.; FERRÉ, J.; BOQUÉ, R. Calculation of the reliability of classification in discriminant partial least-squares binary classification. **Chemometrics and Intelligent Laboratory Systems**, v. 95, p. 122–128, 2009.

PRAWIRA, M.; BARRINGER, S. A. Effects of conching time and ingredients on preference of milk chocolate. **Journal of Food Processing and Preservation**, v. 33, n. 5, p. 571–589, 2009.

REA, A.; REA, W. How many components should be retained from a multivariate time series PCA? *Unpublished results*, 1–49. Avaliable in http://arxiv.org/abs/1610.03588

RENCHER, A. C. **Methods of multivariate analysis.** 2ed. New York: Wiley, 2002. 708 p.

TABOADA, H. A.; COIT, D. W. Data clustering of solutions for multipleobjective system reliability optimization problems. **Quality Technology and Quantitative Management**, v. 4, n. 2, p. 191–210, 2007.

WANG, K.; JIANG, W. High-dimensional process monitoring and fault isolation via variable selection. **Journal of Quality Technology**, v. 41, n. 3, p. 247–258, 2009.

WANG, W.; FAN, J. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. **The Annals of Mathematical Statitsics**, v. 45, n. 3, p. 1342–1374, 2017.

WOLD, S. Cross-validatory estimation of the number of components in factor and principal component models. **Technometrics**, v. 20, n. 4, p. 397–405, 1978.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 109–130, 2001.

WOODALL, W.; MONTGOMERY, D. Some current directions in the theory and application of statistical process monitoring. **Journal of Quality Technology**, v. 46, n. 1, p. 78–94, 2014.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G.J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z-H; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, p. 1–37, 2008.

YAN, Z.; KUANG, T. H.; YAO, Y. Multivariate fault isolation of batch processes via variable selection in partial least squares discriminant analysis. **ISA Transactions**, v. 70, p. 389–399, 2017.

ZARZO, M.; FERRER, A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. **Chemometrics and Intelligent Laboratory Systems**, v. 73, n. 1, p. 15–27, 2004.

ZHAO, C.; WANG, W. Efficient faulty variable selection and parsimonious reconstruction modelling for fault isolation. **Journal of Process Control**, v. 38, p. 31–41, 2016.

ZHAOMIN, L.; QINGCHAO, J.; XUEFENG, Y. Batch process monitoring based on multisubspace multiway principal component analysis and time-series bayesian inference. **Industrial & Engineering Chemistry Research**, v. 53, p. 6457–6466, 2014.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. **IEEE Transactions on Evolutionary Computation**, v. 3, n. 4, p. 257–271, 1999.

5 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as conclusões da tese, além de sugestões para trabalhos futuros.

5.1 CONCLUSÕES

A presente tese teve por objetivo desenvolver um novo método para o controle estatístico de processos industriais em bateladas. A seleção das variáveis mais importantes para maximizar a acurácia de classificação de bateladas foi conduzida visando à construção de um modelo de monitoramento de processo capaz de melhorar o desempenho da detecção de falhas e mitigar as limitações dos métodos tradicionais quando bancos de dados de elevada dimensionalidade, e com duração variável, são analisados. Esse objetivo geral foi alcançado mediante a execução de seis objetivos específicos.

Os dois primeiros objetivos específicos: **identificar as limitações encontradas pelos métodos MSPC no monitoramento de processos industriais**; e **entender como métodos de seleção de variáveis são integrados para promover a melhoria do monitoramento de processos de elevada dimensionalidade** foram alcançados no Artigo 1.

O Artigo 1 apresentou uma revisão sistemática da literatura demonstrando como as limitações dos métodos de MSPC na análise de bancos de dados industriais de elevada dimensionalidade estão sendo solucionadas pela integração a métodos de seleção de variáveis. Assim, foi possível o entendimento do problema de forma a fomentar e justificar a escolha do tema desta tese. Esse primeiro artigo se utilizou de uma análise qualitativa, com o intuito de mapear os métodos publicados que propuseram o uso de seleção de variáveis para promover a melhoria dos métodos de MSPC. A evolução do estado da arte nesse tópico foi demonstrada sendo cada um dos 30 métodos identificados na literatura brevemente descritos e discutidos. O artigo inovou ao propor uma classificação dos métodos propostos em relação a abordagem de seleção de variáveis implementada, e ao categorizar esses em 10 *clusters* de acordo seus objetivos e etapa de monitoramento de processo para a qual foram desenvolvidos. Assim, o artigo contribui para auxiliar pesquisadores no desenvolvimento desse tópico mediante a definição das lacunas existentes sinalizando para oportunidades de pesquisas futuras, bem como auxiliando profissionais responsáveis por departamentos de qualidade a identificar métodos que o possam ajudar na solução de problemas industriais reais.

O terceiro e quarto objetivos declarados, **discutir sobre métodos para alinhamento e sincronização de bateladas aplicados a processos com diferentes durações**; e **definir o método de alinhamento e sincronização mais adequado para o tratamento de dados de bateladas, visando aprimorar a construção do modelo de monitoramento na Fase I do SPC** foram encaminhados no Artigo 2.

O Artigo 2 buscou identificar o tratamento adequado a ser realizado para eliminar a duração variável existente em um banco de dados em bateladas real da conchagem do chocolate ao leite, de forma a permitir sua posterior análise por métodos de MSPC. Seguindo uma abordagem quantitativa, três métodos de DTW foram aplicados para promover o alinhamento e sincronização das trajetórias de 4 variáveis coletadas de 62 bateladas com durações entre 495 e 1.170 minutos. Os resultados foram considerados satisfatórios, sendo exitosa a aplicação dos métodos no preparo do banco de dados analisado. Isto pode ser evidenciado nos resultados obtidos para as trajetórias analisadas, bem como na identificação das fases do processo de conchagem quando os resultados foram avaliados do ponto de vista da tecnologia de fabricação do chocolate. Posterior ao alinhamento, o objetivo é a utilização desse banco de dados alinhado e sincronizado na construção da distribuição de referência para monitoramento do processo em bateladas. Assim, o desempenho de classificação das bateladas em conformes e não conformes foi verificado. Mediante aplicação da técnica de classificação por kNN, o método proposto por Kassidas, MacGregor e Taylor (1998) foi considerado o mais adequado para tratar esse banco de dados já que apresentou a melhor combinação das métricas de desempenho (acurácia, sensibilidade e especificidade), mantendo o poder de classificação das bateladas após as mesmas serem ajustadas para um mesmo tempo de duração, além de ter requerido o menor número de vizinhos (k=3) para obtenção desses resultados. De acordo com esse método, a variável mais importante para o processo de alinhamento e sincronização, e a mais consistente de batelada para batelada, foi a ‘Corrente do motor da concha’. A maioria dos métodos de monitoramento de processos tem seu foco no desempenho da Fase II do monitoramento (PERES; FOGLIATTO, 2018; WOODALL; MONTGOMERY, 2014). Nesse contexto, até onde se tem conhecimento, a análise do impacto dos diferentes métodos de alinhamento e sincronização na determinação do conjunto de referência de bateladas conformes não foi previamente explorada, fazendo com que esse artigo contribua de maneira inovadora para o desenvolvimento de modelos sob-controle mais adequados para a Fase I do CEP.

Por fim, os dois últimos objetivos específicos, **propor a seleção de variáveis, com propósito de classificação, prévia à construção das CCM baseadas em PCA para monitorar um processo em bateladas**; e **validar o desempenho de detecção de falhas da carta de controle multivariada proposta em comparação às cartas tradicionais $\boldsymbol{T^2}$ e $\boldsymbol{Q}$ baseadas em PCA** foram atingidos no Artigo 3.

O Artigo 3 propôs o método PVS-MPCA para detecção de falhas em bancos de dados de elevada dimensionalidade de processos industriais em bateladas. O banco de dados bidimensional com 2.864 variáveis desdobradas e 62 bateladas, pré-tratado, alinhado e sincronizado, foi utilizado para o estudo de caso do método proposto. O número de bateladas conformes e não conformes era conhecido *à priori*, baseado nas análises do laboratório de qualidade da indústria. Sendo assim, a seleção seguiu uma abordagem *wrapper* envolvendo um índice de importância de variáveis, baseado nos parâmetros de saída da Análise Discriminante em Mínimos Quadrados Parciais (PLS-DA ou *Partial Least Squares – Discriminant Analysis*), e a técnica de classificação kNN. Dessa forma, a cada iteração a variável menos importante para a classificação das bateladas foi descartada e a acurácia de classificação do subgrupo remanescente avaliada. Após a implementação da análise de otimalidade de Pareto, o subgrupo com 0,17% de variáveis retidas foi considerado o que maximizava a acurácia em 100% com o menor número de variáveis (5 variáveis desdobradas retidas). Essas variáveis representavam pontos de transição de fases em 3 das 4 variáveis originais coletadas no processo. O desempenho desse subgrupo de variáveis foi comparado ao desempenho do grupo com todas as variáveis na construção do modelo de referência (Fase I) e do modelo de monitoramento *off-line* de bateladas futuras (Fase II). Foi verificado um colapso na matriz de correlações quando a análise PCA foi executada no banco de dados completo compreendido por uma quantidade de variáveis muito superior a quantidade de observações. Na construção do modelo de referência, as cartas T^2 e Q sinalizaram que 27 das 35 bateladas consideradas conformes pela indústria eram não conformes, totalizando em 77,14% de taxa de alarme falso (FAR). Quando essas cartas foram construídas baseadas no subconjunto com 5 variáveis selecionadas, a FAR reduziu em 85,18% sinalizando erroneamente somente 4 das 35 bateladas consideradas conformes pela indústria (FAR = 11,43%). As bateladas futuras (não conformes segundo a indústria) foram corretamente sinalizadas em ambas as situações (Fase II). Tendo em vista que os métodos publicados para detecção e isolamento de falhas em bancos de dados em bateladas de elevada dimensionalidade têm sido propostos para substituir o uso das cartas de controle baseadas em PCA (CHU; QIN; HAN, 2004; YAN; KUANG; YAO, 2017; ZARZO; FERRER, 2004), o método PVS-MPCA surge como uma proposta inovadora para estender o uso das tradicionais CCM T^2 e Q na detecção de falhas, mitigando suas limitações.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Pesquisas futuras podem ser desenvolvidas como extensões dos desenvolvimentos aqui propostos. São elas:

a) Propor um novo método de alinhamento e sincronização de dados em bateladas de duração variável.
b) Avaliar o uso de outros métodos de seleção para identificar as variáveis mais importantes para o monitoramento de processos.
c) Propor um método de seleção de variáveis que não necessite de um grande número de bateladas não conformes para treinamento do algoritmo de classificação.
d) Analisar o impacto promovido pela seleção de variáveis no isolamento de falhas pelos gráficos de contribuição.
e) Validar a implementação do método proposto em diferentes ramos de atuação industriais.
f) Comparar o método proposto com outros métodos de detecção e isolamento de falhas em processos em bateladas de elevada dimensionalidade.
g) Estender o método proposto para monitoramento de processos em bateladas em tempo real.

5.3 REFERÊNCIAS

CHU, Y.-H.; QIN, S. J.; HAN, C. Fault Detection and Operation Mode Identification Based on Pattern Classification with Variable Selection. **Industrial & Engineering Chemistry Research**, v. 43, p. 1701–1710, 2004.

KASSIDAS, A.; MACGREGOR, J. F.; TAYLOR, P. A. Synchronization of batch trajectories using dynamic time warping. **AIChE Journal**, v. 44, n. 4, p. 864–875, 1998.

PERES, F. A. P.; FOGLIATTO, F. S. Variable selection methods in multivariate statistical process control: A systematic literature review. **Computers and Industrial Engineering**, v. 115, p. 603–619, 2018.

WOODALL, W.; MONTGOMERY, D. Some current directions in the theory and application of statistical process monitoring. **Journal of Quality Technology**, v. 46, n. 1, p. 78–94, 2014.

YAN, Z.; KUANG, T. H.; YAO, Y. Multivariate fault isolation of batch processes via variable selection in partial least squares discriminant analysis. **ISA Transactions**, v. 70, p. 389–399, 2017.

ZARZO, M.; FERRER, A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. **Chemometrics and Intelligent Laboratory Systems**, v. 73, p. 15–27, 2004.