Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística

Carta de Controle para Processos em Batelada através de uma Abordagem "Model Free" Utilizando U-estatísticas

Renan Faraon Cintra

Porto Alegre, junho de 2022.

Dissertação submetida por Renan Faraon Cintra como requisito parcial para a obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade Federal do Rio Grande do Sul.

Orientador:

Dr. Danilo Marcondes Filho (PPGEst - UFRGS)

Coorientador:

Dr. Marcio Valk (PPGEst - UFRGS)

Comissão Examinadora:

Dr. Fábio Mariano Bayer (PPGEst - UFRGS)
Dr. Silvana Schneider (PPGEst - UFRGS)
Dr. Ângelo Marcio Oliveira Sant'anna (UFBA)

Data de Apresentação: 02 de Junho de 2022

Agradecimentos

Agradeço:

- A minha mãe Ana Maria, por todo apoio.
- A Fernanda, pela parceria.
- Ao prof. Dr. Danilo Marcondes Filho, meu orientador, por toda ajuda e empenho pra que esse trabalho se realizasse.
- Ao prof. Dr. Marcio Valk, coorientador, por toda dedicação.
- Aos membros da banca, pela disponibilidade.
- Aos colegas, funcionários e professores do Programa de Pós-Graduação em Estatística da UFRGS, pela convivência e amizades.

Resumo

Este trabalho propõe uma abordagem *Model Free* baseada na teoria das U-estatísticas para monitorar processos em batelada. Processos em batelada produzem séries temporais, em que cada uma delas representa sucessivas medições de uma variável de processo, sendo o principal desafio capturar tanto a variabilidade no domínio das bateladas (variabilidade de batelada para batelada) quanto no domínio do tempo (variabilidade serial e correlação cruzada nas variáveis). As abordagens clássicas são focadas no primeiro objetivo, aplicando técnicas nas colunas de uma matriz de dados, na qual cada linha contém os dados de toda a execução de uma batelada. Essas abordagens são fundamentadas na tradicional *Multiway Principal Component Analysis* (MPCA). Por outro lado, recentes abordagens enfatizam o segundo objetivo, levando em conta a natureza temporal dos dados, usando modelos de séries temporais tradicionais como ARMA (*Autoregressive Moving Average*)/VAR (*Vector Autoregressive*) para modelar diretamente a variabilidade no domínio do tempo. Através da teoria das U-estatísticas, propõe-se construir um conjunto de Cartas de Controle capazes de monitorar ambas as fontes de variabilidade conjuntamente. Adicionalmente, o monitoramento baseado na U-estatística, ora proposto, tem como importante característica sua flexibilidade, pois evita-se a necessidade de identificação de modelos, estimação de parâmetros e acomoda as duas fontes de variabilidade mesmo num contexto de poucas bateladas e muitos instantes de tempo. Além disso, permite um monitoramento de diferentes estruturas de dados temporais isoladamente, sendo assim bastante adequada no contexto de processos em batelada. Através de experimentos numéricos em dados simulados e reais, mostrou-se que a abordagem baseada na U-estatística apresenta bom desempenho em diferentes cenários.

Índice

1 Introdução

Uma das ferramentas mais importantes no Controle Estatístico do Processo (CEP) são as Cartas de Controle (CC). As CCs apresentam um resumo visual (gráficos) para indicar se um processo está em estado que satisfaz certos critérios. As CCs são construídas a partir de amostras históricas utilizando a distribuição amostral de alguma característica de interesse (Média, Mediana, Desvio Padrão,...) dado que o processo está sob causas comuns de variação (ou sob controle), sendo seus limites de controle obtidos a partir dos quantis da distribuição definidos a partir de uma probabilidade de alarme falso α. As CCs dessa forma servem como ferramenta para diagnosticar a qualidade de uma certa produção.

No contexto de processos em bateladas, tem-se dados históricos representando séries temporais de medições de variáveis ao longo da batelada (variáveis de processo). O desafio é construir CCs para monitorar novas bateladas verificando se o comportamento das novas séries temporais está de acordo com o padrão das séries de referência. Bateladas com variáveis apresentando trajetórias distante das esperadas indiciam que de alguma forma o processo está com causas de variabilidade não inerentes do processo, indicando que o produto final pode não estar de acordo com suas especificações.

Processos em bateladas é um tema relevante no CEP devido sobretudo à sua peculiar estrutura de dados. Presentes em diversos setores da indústria, como a química, alimentícia e farmacêutica, os processos em bateladas têm a característica de serem tridimensionais - I-bateladas, K-variáveis e T-instantes no tempo -, o que torna mais complexo elaborar uma abordagem para o monitoramento destes processos. Nesse sentido, o desafio é propor CCs que levem em conta tanto a variabilidade entre bateladas como a proveniente da estrutura temporal (correlação serial dos dados).

A literatura de CEP orientada a esses processos é marcada pelo trabalho inaugural de Nomikos and MacGregor [1995] e se centraliza em cartas de controle em que a estrutura de dados é rearranjada numa disposição bidimensional. Esse tipo de abordagem, que se utiliza do método multivariado *Multiway Principal Component Analysis* (MPCA), incorre numa perda da captura das caraterísticas temporais dos processos. Por outro lado, mais recentes trabalhos têm focado em tratar dessa dimensão temporal, se utilizando de modelos clássicos como ARMA e VAR (Choi et al. [2008], Pan and Jarrett [2012], Vanhatalo and Kulahci [2015], Marcondes Filho and Valk [2020] de Oliveira et al. [2022]).

No que se refere à literatura Estatística a respeito de métodos de classificação, existem inúmeras técnicas, muitas delas atualmente baseadas em *machine learning*. Especificamente, dado o contexto de monitoramento de processos em batelada, o problema de classificação aqui pode ser relacionado a uma abordagem conhecida na literatura como *one-class-classification*, que abrange as técnicas de detecção de *outliers* e detecção de novidades (Pimentel et al. [2014], Chandola et al. [2009]), que, geralmente, sofrem limitações severas nas propriedades inferenciais quando utilizadas em conjunto de dados caracterizados por grandes dimensões e poucas observações (HDLSS -*High-dimension low-sample* *size*), algo comum no contexto de CEP de processos em bateladas.

A ideia a ser desenvolvida tem como base uma série de trabalhos que se incia com Sen [2006], que propõe uma solução para um teste de comparação de médias (MANOVA) num contexto genômico onde a dimensão das variáveis apresentada é maior que a quantidade de observações. De maneira geral, o autor usa medidas de distâncias (*Hamming-distance*) entre grupos e dentro de grupos, ponderadas pelo tamanho dos grupos de forma estratégica.

Posteriormente, Pinheiro et al. [2009] mostram que a estatística apresentada no contexto específico genômico de fato pode ser usada de forma mais geral, apresentando propriedades como normalidade assintótica sem assumir homogeneidade ou normalidade dos dados. Utilizando o fato de que a estatística proposta por Sen [2006] é uma U-estatística, Pinheiro et al. [2009] mostram que, sob a hipótese de que todas as observações vêm de uma mesma distribuição, a normalidade assintótica é válida mesmo em contextos em que os dados tenham dependência na dimensão (exemplo: séries temporais). Além disso, a convergência ocorre com aumento da dimensão e/ou com o aumento do tamanho amostral.

Valk and Pinheiro [2012] estudam o comportamento dessa U-estatística no contexto de séries temporais, apresentando um método de agrupamento com inferência, ou seja, os grupos resultantes seriam associados a uma significância estatística. No entanto, a abordagem proposta depende de testar todas as combinações possíveis para determinar dois grupos, o que é computacionalmente custoso. Cybis et al. [2018] encontram o agrupamento que maximiza a U-estatística de teste, através de um processo de otimização. Para esse autores, a partir da distribuição do máximo de várias U-estatísticas, é possível inferir se a separação obtida dos dados em dois grupos caracterizaria de fato grupos estatisticamente separados. Isso permitiria concluir pela homogeneidade entre dois grupos, caso não tenha sido detectada uma diferença significativa na configuração que resultou na maior separação possível.

Porém, uma restrição dessa teoria é a necessidade de grupos com pelo menos duas observações. Valk and Cybis [2020] trazem uma extensão da estatística de teste, permitindo que um dos grupos apresente tamanho $n = 1$. Esse desdobramento tornou possível encontrar uma separação de n observações em dois grupos em que um deles tenha $n - 1$ observações e o outro $n = 1$, e, finalmente, testar se essa separação é estatisticamente significativa, uma abordagem propícia ao contexto de detecção de *outlier* ou *one-class-classification*.

O método ora proposto utilizará a estatística estendida por Valk and Cybis [2020] com o objetivo de verificar se uma nova amostra tem a mesma distribuição daquelas do grupo de amostras de referência, com base nas medidas de distância entre as características de interesse dos dados. Dessa forma, em um contexto de processo em bateladas, considerando cada amostra como a série temporal de variáveis do processo, será utilizada a distribuição da estatística desenvolvida por Valk and Cybis [2020] para avaliar se amostras de determinados lotes estariam de acordo com os lotes de referência.

Este trabalho está organizado da seguinte forma: nesta seção, apresentou-se uma contextualização do tema; na Seção 2, descreve-se os objetivos, geral e específicos; a Seção 3 é dedicada a apresentar aspectos gerais das abordagens de processos em batelada na literatura de Controle Estatístico; uma breve introdução à teoria das U-Estatísticas e como ela será utilizada neste trabalho no contexto de Cartas de Controle é desenvolvida na Seção 4; Seção 5 resume a abordagem aqui proposta; Seção 6, o artigo da dissertação e, finalmente, Seção 7, a Conclusão.

2 Objetivo

2.1 Objetivo Geral

Desenvolver um conjunto de cartas de controle para processos em bateladas, considerando as duas fontes fundamentais de variabilidade de tais processos baseadas na teoria de U-estatísticas.

2.2 Objetivos Específicos

- Apresentar a abordagem proposta para monitoramento de tais processos baseada na teoria de U-estatísticas.
- Realizar estudo com dados simulados comparando o desempenho da abordagem proposta com abordagem tradicional.
- Verificar o desempenho da abordagem proposta em dados de processos reais.

3 CEP Tradicional para Processos em Bateladas

Sendo processos amplamente utilizados, os processos em bateladas são caracterizados por uma estrutura de dados tri-dimensional: I bateladas × K variáveis × T instantes no tempo do processo. Dado um contexto de processos químicos por exemplo, um processo em bateladas pode ser descrito por três etapas (Marcondes Filho [2001]), em que, primeiro, determinados componentes (matérias primas) são colocados dentro de um recipiente; posteriormente, estes componentes sofrem alterações (transformações químicas) devido a algum estímulo. Durante esta etapa, inúmeras variáveis de processo são sucessivamente medidas a cada instante de tempo; finalmente, tem-se o produto final (no caso do processo químico, o produto deve ser um líquido), que é analisado segundo especificações. Após esta etapa, retoma-se à primeira, de modo que se inicia uma nova batelada. O monitoramento de uma nova batelada deve levar em conta duas fontes de variabilidade (Ge et al. [2013]): a variabilidade entre bateladas (variabilidade entre as séries temporais da mesma variável em diferentes bateladas) e dentro das bateladas (correlação serial nos dados das séries temporais de cada variável dentro de uma batelada).

Tendo em vista essa especificidade, a discussão da literatura sobre processos em batelas se centraliza na proposição de cartas de controle baseadas em técnicas de estatística multivariada na estrutura de dados *three-way* desdobrada numa estrutura *two-way* de duas formas: (i) O arranjo ($I \times KT$) proposto no trabalho de Nomikos and MacGregor [1994]. Neste arranjo, a correlação serial entre as variáveis é capturada não diretamente via modelagem clássica de séries temporais onde nas colunas as medições das diferentes variáveis em diferentes instantes de tempos estão empilhadas nas linhas. A Figura 1 (a) ilustra tal arranjo. (ii) O arranjo ($IT \times K$) foi proposto no trabalho inicial de Wold et al. [1998]. Aqui a correlação entre variáveis (cruzada) é priorizada. A Figura 1

(b) destaca este caso. Ambos os métodos, por suas estruturações de arranjo nos dados, acabam priorizando a variabilidade entre bateladas em detrimento da variabilidade dentro das bateladas (correlação serial dos dados das variáveis de processo).

Abordagens mais recentes baseadas em modelos de séries temporais por definição priorizam a modelagem da correlação serial através do uso principalmente de modelos da classe ARMA (*Autoregressive Moving Average*) e VAR (*Vector Autoregressive*). Trabalho precursor nesta direção pode ser encontrado em Choi et al. [2008]. Atento ao problema da modelagem de dados em bateladas considerando as duas fontes de variabilidade simultaneamente, algumas abordagens atuais propõem modificações em modelos de séries temporais para acomodar a variabilidade entre bateladas, visto que a teoria inferencial de séries temporais está construída a partir de uma uma série temporal (modela a variabilidade dentro da série ou modela a correlação serial), isto é, de uma batelada apenas. Destacamos o trabalho nesta direção de Wang et al. [2017].

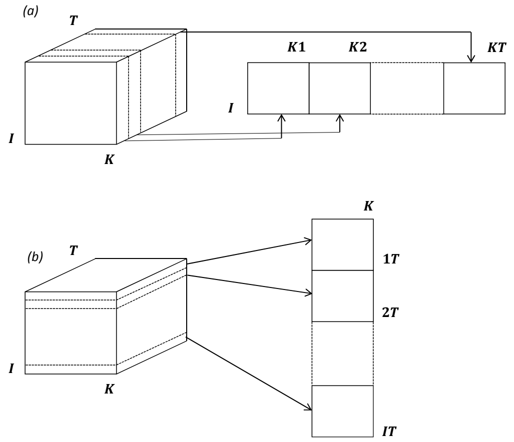

Figura 1: (a) Arranjo no domínio do tempo: $I \times KT$ proposto por Nomikos e Macgregor (1994). (b) Arranjo no domínio das variáveis : $IT \times K$ proposto por Wold et al. (1998).

Esta dissertação propõe uma abordagem alternativa que considera ambas as fontes de variabilidade para controle estatístico de processos em bateladas. O método do trabalho ora proposto utiliza U-estatísticas apresentadas por Valk and Cybis [2020] e tem a vantagem de ser uma abordagem *Model Free*. A ideia é usar o método de classificação com U-estatísticas para determinar se um conjunto de bateladas pertence a um grupo de bateladas sob controle, considerando a dimensão temporal do processo. Assim, através de condições bastante flexíveis, propõe-se um método que permite diagnosticar mudanças estruturais da dinâmica dos dados, tais como mudança na correlação serial das variáveis de processo e na tendência das séries temporais (incluindo mudanças na média das séries temporais) .

4 Cartas de controle e U-estatísticas

4.1 Uma breve introdução a U-Estatísticas

U-estatística é uma classe de estatísticas com destacada relevância na teoria de estimação, especialmente na construção de estimadores não viesados de variância uniformemente mínima (ENVVUM). A teoria clássica de U-estatística foi introduzida por Hoeffding [1948], mas outras referências podem ser consultadas, como Lee [1990], Fraser [1956] capítulo 6,Denker [1985] e Lehmann [1999]. Alguns estimadores conhecidos na literatura pertencem à classe das U-estatísticas, por exemplo, momentos de uma distribuição (média e variância), estatística não paramétrica do teste de Wilcoxon, *Testing Symmetry* e Medidas de Associação.

De modo geral, uma U-estatística é baseada em uma amostra aleatória independente e identicamente distribuída (iid) $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ de uma função de distribuição $\mathbf{F}$ e um parâmetro tal que exista a seguinte relação:

$$\theta = \mathbb{E}\left[\phi\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_\zeta\right)\right], \tag{1}$$

onde $\zeta \leq n$ e ϕ é uma função simétrica. Então, define-se a U-estatística por:

$$U_n = \binom{n}{\zeta}^{-1} \sum_{C_{n,\zeta}} \phi\left(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_\zeta}\right), \tag{2}$$

onde $C_{n,\zeta}$ representa todas combinações de ζ elementos em n.

No caso deste trabalho, a abordagem proposta se baseia numa estatística que pertence à classe das U-estatísticas e, portanto, possui as seguintes vantagens: sua distribuição assintótica não depende da suposição de normalidade dos dados nem de homocedasticidade; nas variáveis, é permitido alguma estrutura de dependência; e a convergência assintótica ocorre com o aumento do número de variáveis (dimensão), o que é especialmente favorável em contexto de *HDLSS.*

4.2 A U-Estatística do Método

Considere $\mathbf{X}_1 \dots \mathbf{X}_I$ uma amostra de vetores T-dimensional e que cada $\mathbf{X}_i$ seja um processo pertencente a um grupo de processos sob (G_1) ou fora (G_2) de controle, com tamanho amostral respectivamente n_1 e n_2, onde $I = n_1 + n_2$. Além disso, assume-se que para cada grupo, com $g \in 1, 2$, $\mathbf{X}_1^{(g)} \dots \mathbf{X}_{n_g}^{(g)}$ são independentes e identicamente distribuídas com distribuição F_g com vetor de médias μ_g e matriz de dispersão $\mathbf{\Sigma}_g$ positiva definida, com g .

A partir de Sen [2006] e Pinheiro et al. [2009], define-se a seguinte função de distância entre os grupos $\theta(F_g, F_{g'})$

$$\theta(F_g, F_{g'}) = \int\int \phi(x_1, x_2) dF_g(x_1) dF_{g'}(x_2) \quad , \quad x_1, x_2 \in \mathbb{R}^{\mathbb{T}} \tag{3}$$

onde g, $g' \in \{1, 2\}$ e $\phi(\cdot, \cdot)$ é um kernel simétrico de segunda ordem. Assumindo $\phi(\cdot, \cdot)$ como uma função convexa linear de seus componentes marginais, tem-se

$$\theta(F_g, F_{g'}) \geq \frac{1}{2}\{\theta(F_g, F_g) + \theta(F_{g'}, F_{g'})\}, \tag{4}$$

para todas as distribuições de F_g e $F_{g'}$, cuja igualidade mantém-se quando $\mu_g = \mu_{g'}$ (em caso de homoscedasticidade).

A função $\theta(\cdot, \cdot)$ pode ser usada para medir tanto distância entre grupos como dentro de cada grupo. A partir de Hoeffding [1948], segue que um estimador não viesado para função de distância intra-grupo $\theta(F_g, F_g)$ é uma U-estatística generalizada, com kernel $\phi(\cdot, \cdot)$ definida por

$$U_{n_g}(g) = \binom{n_g}{2}^{-1} \sum_{1 \leq i \leq j \leq n_g} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g)}), \tag{5}$$

onde $g \in \{1, 2\}$. E entre grupos

$$U_{n_g,n_{g'}}^{(g,g')} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}), \tag{6}$$

com $g, g' \in \{1, 2\}$ e $g \neq g'$. Quando assume-se que não existem grupos, ou seja, os dados são homogêneos, a U-estatística geral

$$U_n = \binom{I}{2}^{-1} \sum_{1 \leq i \leq j \leq I} \phi(\mathbf{X}_i, \mathbf{X}_j),$$

pode ser decomposta da seguinte forma

$$\begin{aligned} U_n &= \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \sum_{1 \leq g < g' \leq 2} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g,n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\} \\ &= W_n^* + B_n^*. \end{aligned} \tag{7}$$

Onde $W_n^* = \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)}$ e $B_n^* = \sum_{1 \leq g < g' \leq 2} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g,n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\}$. As propriedades estatísticas da B_n^* são discutidas em Pinheiro et al. [2009]. No entanto, fica claro, a partir da definição, que os grupos precisam ser de tamanho mínimo 2. Valk and Cybis [2020] propõem uma extensão que permite que se utilize essa metodologia para casos em que um dos grupos tem tamanho 1. A estatística B_n, estatística central para a metodologia aqui apresentada, é o resultado desse desenvolvimento:

$$B_n = \frac{1}{n} \left(U_{1,n-1}^{(1,2)} - U_{n-1}^{(2)} \right). \tag{8}$$

em que $U_{n-1}^{(2)}$ é a U-estatística associada à distância dentro do grupo, como definido por (5), e $U_{1,n-1}^{(1,2)}$ é a U-estatística referente à medida entre grupos, definida por (6).

Tal estatística é central para o trabalho ora proposto, pois, como demonstra Valk and Cybis [2020], possui relevantes propriedades, tais como, sob a hipótese de homogeneidade entre grupos, ter distribuição assintótica normal. Assim, é possível utilizar os quantis da distribuição para, no caso de uma nova amostra, decidir se esta apresenta mesma distribuição que a de uma amostra homogênea determinada. No contexto de monitoramento de processos em batelada, pode ser usada para determinar se um lote de bateladas provém de um processo de referência, sob controle. Isso será feito basicamente a partir de um teste de hipóteses, onde a estatística do teste, denominada V, será a estatística B_n padronizada.

5 Resumo da Abordagem Proposta

Neste trabalho será apresentada uma abordagem para monitoramento de processos em bateladas através da proposição de cartas de controle baseadas na teoria de U-Estatísticas. A partir de uma estatística para classificação derivada dessa teoria desenvolvida por Valk and Cybis [2020], esta dissertação apresenta uma regra de classificação que permite comparar uma nova batelada em relação às bateladas históricas do processo sob controle. Como cada batelada gera séries temporais de variáveis de processo, as cartas de controle propostas a partir da teoria de U-Estatísticas permitem monitorar diferentes características das séries. Neste trabalho focamos nas correlações seriais e nas mudanças na média das séries, incluindo mudança na tendência.

6 Artigo

Autores: Renan Faraon Cintra, Marcio Valk, Danilo Marcondes Filho

Título: A model free-based control chart for batch process using U-statistics

Ano: 2022

A model free-based control chart for batch process using U-statistics

Cintra, F.R., Valk, M. and Marcondes Filho, D.

Abstract

This paper proposes a free model approach based on U-statistics theory for monitoring batch processes. Through this theory we can build a group of control charts capable of monitoring this kind of process considering either the variability in the batch domain (batch-to-batch variability) and in the time domain (data with serial correlation). These types of processes are known to generate time series of successive measurements of many process variables in each run and the main challenge is to capture and accommodate the both sources of variability. Classical approaches are focused on the first goal by applying multivariate techniques in the columns of a data matrix **X**, in which each row has the data for the entire batch run. Those approaches are grounded in a traditional Multiway Principal Component Analysis (MPCA). Recent approaches are focused on the second goal, taking into account the time series nature of the data by using time series models like ARMA/VAR to directly model the variability in a time domain. The U-based control proposed approach seems to be really flexible since we avoid model identification, parameter estimates issues and it is able to accommodate the between and within batch variability. Additionally, the proposed approach is suitable to model a wide range of batch data feature, like drifts (including trends) and serial correlation (data dynamics). Unlike in the MPCA-based approaches, our proposition is able to accommodate the size of data matrix **X** for the large number of columns (i.e., a large number of variables × time-instants). Through the simulated and real data we show that the U-based approach works well in a wide range of batch processes, including a scenario with a very low number of reference batches available.

1. Introduction

A vast range of items are produced by using industrial batch operations. This kind of process generates for each batch a time-series of successive measures of each process variable, resulting in a three-way data structure (batches × variables × time-instants). The main goal is monitoring future batches considering the batch-to-batch variability and the data serial correlation (inner-batch variability). Classical monitoring approaches do not fully take into account the time series nature of the data. Most of them decompose the data in the two-way array (batches × variables/time-instants), so that each row represent the overall information about one batch including the successive measurements of the process variables. In this context, considering batches as sample replications, control approaches are proposed by using multivariate techniques applied in the variables × time-instants columns of that two-way array. These multivariate-based control charts are capable of capturing directly the batch-to-batch variability and the data serial correlation in some way. A ground theory in this direction is in the precursor work of [1]. A good view of those approaches including improvements and applications can be found in ,[2], [3], [4], [5], [6], [7],[8], [9] and [10].

Another group of approaches look for the balance between the batch-to-batch and the inner-batch variability, by doing a bunch of alternative unfolding of the tree-way batch data. They try to mix the information of each batch in the rows and columns of a two-way array and model those data applying multivariate techniques on that. Some work in that direction can be found in [11], [12],[13], [14], [15], [16], [17], [18], [19], [20] , [21] and [22].

There are some alternative approaches that consider directly the time series nature of batch data. They seek for the balance between the two sources of variability (like the previous approaches) assuming that the batches are generated from one stochastic process. In short, in the first step, for each batch, the data serial correlation are modeled by the traditional ARMA (*Autoregressive Movind Average*) and VAR (Vector Autoregressive) time series models (inner-batch variability). In the second step some statistics are proposed to accommodate the coefficient estimates for each batch (batch-to-batch variability). Works in this direction can be found in [23], [24] ,[25] and most recently in [26].

This paper proposes a new control approach based on the U-statistics theory of [27]. It is a completely free model and free data distributed approach. The well defined U-based V statistic is aimed to classify if the new samples belong to a group of reference samples, based on the distance measures between input data vectors from those samples. In a batch process context, considering each sample as the time series of process variables, we propose the V statistic to evaluate batch samples according to the reference batches. A wide range of time series data functions can be used as the input vector in this statistic as well, so that we can consider different common features of time series batch data, including drifts (including trends) and data dynamics (data serial correlation). Under the hypotheses of homogeneity between groups and other flexible conditions, the V is asymptotically Normally distributed in T, where T is the size of the batch data (time series length), no matter the number of the reference batches available. It is extremely desirable, since modern data sensors generate a bunch of data in each batch run and the classical approaches are not able to deal with number of time-instants bigger than the number of reference batches [10]. Another point is that, by its structure, V statistic considers both sources of batch variability (between an inner), i. e., the length of the time series and the number of reference batches are well accommodated. We set a group of

V-based control charts to evaluate new batch samples considering drifts and data dynamics. The flexibility and power of those charts are illustrated by using simulated and real data. A comparative study with the fundamental approach of [1] is presented as well.

The paper is organized as follows: Section 2 describes as the benchmark the fundamental approach of [1]. Section 3 brings a detailed description of our methodology. In Section 3, the proposed approach is illustrated through simulated data. Section 4 shows an application involving a real data set. Additional issues are discussed in Section 5 and Conclusions are presented in Section 6.

2. MPCA-based approach

We present as a benchmark approach the well-known classical Multiway Principal Component Analysis method (MPCA) developed by [1], which is briefly described in this section. Assuming that I batches from in-control process are available, each one with K process variables measured during T time-instants. In order to modeling and monitoring the data correlation, [1] consider the data arranged in a two-way matrix $\boldsymbol{X}$, with dimension $(I \times KT)$. They propose two control charts based in the Principal Component Analysis (PCA) applied in the columns of $\boldsymbol{X}$, so that each variable observed at a given time instant is treated as a new variable.

The PCA consists in diagonalize the sample variance-covariance matrix $\boldsymbol{S}_X$ obtained from $\boldsymbol{X}$ and save the sample eigenpairs $(\lambda_j, \boldsymbol{u}_j)$, for $j = 1, ..., KT$. The eigenvectors $\boldsymbol{u}_j$, with dimension $KT \times 1$, represent the linear combination of the KT variables that transforms them in new non-correlated variables [named as Principal Components (PC's)] and the eigenvalues λ_j is the variance of the associated PC's.

Since the KT variables in X are highly correlated, a few vectors $\boldsymbol{u}_j$ can give enough description of the overall variability in $\boldsymbol{X}$, i. e., a number $L(< KT)$ of PC's from the L largest eigenvalues λ_j are needed. The score of the j^{th} retained PC is obtained by $y_j = \boldsymbol{x}\boldsymbol{u}_j$, where $\boldsymbol{x}$ is the row vector (with dimension $1 \times KT$) with the overall information of one batch arranged according to the $\boldsymbol{X}$.

The new batch sampled is monitoring through the L retained PC's by using the $\mathcal{T}^2_{pca}$ Hotelling statistic and the PCA based-residual $\mathcal{Q}$ statistic. Since the PC's are uncorrelated and the data in $\boldsymbol{X}$ is centered, the $\mathcal{T}^2_{pca}$ quantity can be written as the sum of L terms like ([28]):

$$\mathcal{T}^2_{pca} = \sum_{j=1}^{L} \frac{y_j^2}{\lambda_j} \dot{\sim} \frac{(I-1)L}{(I-p)} F_{L,I-L}, \tag{1}$$

where $\dot{\sim}$ means asymptotic convergence in distribution, when I increases, and $F_{L,I-L}$ is the Snedcor Distribution. The quantity $\mathcal{Q}$ has the following form ([28]):

$$\mathcal{Q} = \sum_{j=L+1}^{KT} \frac{y_j^2}{\lambda_j} \dot{\sim} A\left[z \times B + 1 + \mathcal{C}\right]^{1/h}, \tag{2}$$

where $\dot{\sim}$ means asymptotic convergence in distribution, when I increases, and z is the quantile from the standard Normal Distribution. $A = \sum_{j=L+1}^{KT} \lambda_j$, $B = \sqrt{2\sum_{j=L+1}^{KT} \lambda_j^2 h^2}/A$, $\mathcal{C} = \sum_{j=L+1}^{KT} \lambda_j^2 h(h-1)/(\sum_{j=L+1}^{KT} \lambda_j)^2$ and $h = -2[(\sum_{j=L+1}^{KT} \lambda_j)(\sum_{j=L+1}^{KT} \lambda_j^3)]/3(\sum_{j=V+1}^{KT} \lambda_j^2)^2$ are obtained from the $(KT-L)$ non retained eigenvalues λ_j.

The MPCA-based control charts are built using these two quantities. The $\mathcal{T}_{pca}^2$ chart monitors the behavior of known sources of process variability, i.e., deviations of the time trajectories of the variables from their reference trajectories in $\boldsymbol{X}$. The $\mathcal{Q}$ chart is used to detect unusual events that affect the cross and serial correlation of the reference data in $\boldsymbol{X}$, captured by the retained PC's.

Follow the aim of our approach that considers one variable at a time, we assume here $K = 1$, so that each row of the reference data in $\boldsymbol{X}$, with dimension $I \times T$, represents a time series describing the trajectory of one process variable in the i^{th} batch, for $i = 1, ..., I$.

3. U-Statistics based approach

In this section, the ground of our methodology is described, including the description of the statistics derived from U-statistic that define a measure of variability within and between groups. We are particularly interested in one of these statistics, its properties and the asymptotic distribution. The proposed control approach (derived from that statistic) is described in detail.

3.1. Within and Between Group Variability based on U-Statistics

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are T-dimensional batch samples, each one bringing the time series representing multiple measurements of one process variable. Consider those n samples split in two groups. Lets assign to the group G_1 the samples considered as the *reference samples*, i. e., samples coming from the in-control process. The G_2 group includes the remaining samples that are considered as the *new samples* to be monitored. The sample sizes of G_1 and G_2 are n_1 and n_2, respectively, where $n = n_1 + n_2$.

In the g-th group, for $g \in \{1, 2\}$, batch samples $\mathbf{X}_1^{(g)}, \ldots, \mathbf{X}_{n_g}^{(g)}$ are assumed to be independent and identically distributed with a T-variate distribution F_g. Here, the distribution F_g admits finite mean vector $\boldsymbol{\mu}_g$ and positive definite dispersion matrix $\boldsymbol{\Sigma}_g$ (not necessarily with multivariate normal distribution). Following the traditional approach of [29] and [30], we define the functional distance $\theta(F_g, F_{g'})$ as

$$\theta(F_g, F_{g'}) = \int\int \phi(x_1, x_2) dF_g(x_1) dF_{g'}(x_2), \quad x_1, x_2 \in \mathbb{R}^T, \tag{3}$$

where $g, g' \in \{1, 2\}$ and $\phi(\cdot, \cdot)$ is a symmetric kernel of order 2. The θ parameter represents the distance between the distribution F_1 of the reference group G_1 and the new samples distribution F_2.

If we assume that $\theta(\cdot, \cdot)$ is a convex linear function of its marginal components, then we have

$$\theta(F_g, F_{g'}) \geq \frac{1}{2} \{\theta(F_g, F_g) + \theta(F_{g'}, F_{g'})\}, \tag{4}$$

for all distributions F_g and $F_{g'}$, with equality holding whenever $\mu_g = \mu_{g'}$ and $\mathbf{\Sigma}_g = \mathbf{\Sigma}_{g'}$.

Note that the functional $\theta(\cdot, \cdot)$ can be used to define both distance within and between groups. It follows from U-statistics theory that an unbiased estimator of this functional for within group distance $\theta(F_g, F_g)$ is a generalized U-statistic [31], with kernel $\phi(\cdot, \cdot)$, defined as

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g)}), \tag{5}$$

where $g \in \{1, 2\}$. Analogously, the unbiased estimator for the between group functional distance $\theta(F_g, F_{g'})$ (i.e., the unbiased estimator for the functional distance between the reference group distribution F_1 and the new samples distribution F_2), is defined by

$$U_{n_g, n_{g'}}^{(g,g')} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}), \tag{6}$$

where $g, g' \in \{1, 2\}$ and $g \neq g'$. Considering the n batch samples as a single group, we can define the combined U-statistic as

$$U_n \quad = \quad \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \phi(\mathbf{X}_i, \mathbf{X}_j). \tag{7}$$

Considering equations from (4) to (7), the U_n in (7) can be decomposed as the linear combination of the two following U-statistics

$$\begin{aligned} U_n &= \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \sum_{1 \leq g < g' \leq 2} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\} \\ &= W_n^* + B_n^*, \end{aligned} \tag{8}$$

where $W_n^* = \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)}$ and $B_n^* = \sum_{1 \leq g < g' \leq 2} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\}$. Such decomposition is very similar to the overall variance decomposition in the ANOVA technique. The W_n^* can be seen as the overall within group measure and the B_n^* statistic is the between group overall measure (the same role as the ANOVA F test). The statistic B_n^* plays the central role of our proposition.

Through this statistic, the decision rule will be proposed in order to evaluate if a new batch sample belongs to the in-control process (G_1 group, in which the reference batch samples were obtained). However, the B_n^* properties holds considering the number of at least two samples in each group (reference G_1 and new samples groups G_2) and all the samples in group G_2 must have the same distribution F_2, in which is not really reasonable in the real batch process context. It makes the use of a control chart based on that statistic unfeasible, since we can not assume that each disturbed batch samples coming from the same distribution (F_2). In the next Section the version of that statistic feasible to our goal is presented. Through this alternative B_n^* we can consider the G_2 group with samples coming from different distributions.

An extension allowing groups of size one was presented in [27], which is given by

$$\begin{aligned} U_n &= \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} \phi(\mathbf{X}_i, \mathbf{X}_j) \\ &= \frac{n-1}{n} U_{n-1}^{(2)} + \frac{1}{n} U_{1,n-1}^{(1,2)} + \frac{1}{n}\left(U_{1,n-1}^{(1,2)} - U_{n-1}^{(2)}\right) \\ &= W_n + B_n, \end{aligned} \tag{9}$$

where $U_{1,n-1}^{(1,2)}$ and $U_{n-1}^{(2)}$ are as defined in (5) and (6), respectively. $W_n = \frac{n-1}{n} U_{n-1}^{(2)} + \frac{1}{n} U_{1,n-1}^{(1,2)}$. Through the B_n statistic explicitly written as

$$B_n = \frac{1}{n}\left(U_{1,n-1}^{(1,2)} - U_{n-1}^{(2)}\right) \tag{10}$$

we can now consider individual new samples as a hole G_2 group, i. e., the new samples can be compared to the reference group samples in G_1.

[27] derived some important properties of that B_n statistic. Let $G_1 = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n-1}\}$ and $G_2 = \{\mathbf{X}_n\}$. Under the null hypothesis of overall group homogeneity, i.e, $H_0 : F_1 = F_2$, we have $\mathbb{E}[B_n] = 0$. Under the alternative hypotheses ($H_1 : F_1 \neq F_2$), mild conditions are required to guarantee that $\mathbb{E}[B_n] > 0$. Additionally, under H_0, the statistics B_n is asymptotically normal distributed in T and the convergence rate is of the $\sqrt{T}$ traditional order. Another important issue is that the theoretical B_n variance does not have the closed form, since we do not want to assume any specific groups distribution F_1 and F_2 and kernel $\phi(\cdot, \cdot)$ in (3). It makes the usage of this statistic really flexible and extremely suitable to be used as the role model for our application as a control rule. In this case, the estimated B_n variance is obtained by using re-sampling procedures (see detailed description in [27]).

3.2. U-statistics-based Control Approach

Consider a historical data set of n_1 batches yielding products compliant with specifications, so that they are assign to the reference group G_1, in which each batch sample is coming from F_1 distribution. For each batch we have a time series representing the trajectory of one variable, measured at T time-instants, from the process under normal regime (in-control sample batches). Lets assume now the ongoing process generating new batch samples, each new batch having an unknown F_2 distribution. The new samples are evaluated (one at a time) according to the G_1 group, so that under $H_0 : F_1 = F_2$ and the B_n distribution is written as ([27])

$$V = \frac{B_n}{\sqrt{\operatorname{Var} B_n}} \dot{\sim} N(0,1), \tag{11}$$

where $\dot{\sim}$ means asymptotic convergence in distribution, when T increases. $\operatorname{Var} B_n$ is obtained by re-sampling procedures (see details in [27]).

The control chart based on the V statistic can be built in order to monitor the behavior of each new batch sample according to the reference group samples. Scores above the α percentile in (11) imply that a new batch are in some way far different from the in-control group. Its important to notice that we reject H_0 for high V scores, since under H_1, $\mathbb{E}(V) > 0$. That´s why we have the one tailed test considering only the upper quantile from the significance level α.

We can use a wide range of functions as to play the role of $\phi(\cdot,\cdot)$ in (3). It opens the applicability of the proposed V-based control chart, since through the proper choice of $\phi(\cdot,\cdot)$ a specific feature of the data can be monitored as discussed in the following. Lets remember that in our case, each batch sample represents a time series sample data, so that through the V chart we can monitor new samples considering different time-series features, including dynamic (serial correlation), trends (including drifts), seasonality, cycles, etc. In a batch process context, the focus is on the data dynamic and trends, as we can find in the time series data coming from a typical process.

3.2.1. Monitoring Drifts and Trends

In order to make the V-based chart capable of monitoring changes the time series that represents data drifts or differences in the data trends, we define as $\phi(\cdot,\cdot)$ the measure based on the Euclidean distance. Lets consider the sampled time series X_T (from the reference G_1 group) and Y_T the new sampled time series. The Euclidean distance d is defined as ([32])

$$d\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \left(\sum_{t=1}^{T} \left(X_t - Y_t\right)^2\right)^{1/2}. \tag{12}$$

Through d we can compare these time-series based on the closeness of their values at specific points of time. That's why this metric is very sensitive to point out differences in the time series mean level (representing either data drifts or trend changes). Additionally, changes in scale are detected as well by using d. The V-based control rule in (11) built by using the $\phi(\cdot,\cdot) = d(\cdot,\cdot)$ distance will be called V_d.

3.2.2. Monitoring Serial Correlation (Data dynamics)

The V-based chart aimed to monitor data serial correlation (data dynamics) are set by using distance metrics different from d. It happens because d distance treat each corresponding points as if they are independent in time, so that V_d is not able to detect changes in data dynamics. For that goal, another group of metric based on the Auto-correlation Function (ACF) and Periodogram functions are proposed. The metric based on the ACF distance can be described as [32]

$$d_{ACF}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \sqrt{\sum_{i=1}^{\ell} \left(\hat{\rho}_{i,X_T} - \hat{\rho}_{i,Y_T}\right)^2}, \tag{13}$$

where $\hat{\boldsymbol{\rho}}_{X_T} = \left(\hat{\rho}_{1,X_T}, .., \hat{\rho}_{\ell,X_T}\right)^\top$ and $\hat{\rho}_{Y_T} = \left(\hat{\rho}_{1,Y_T}, .., \hat{\rho}_{\ell,Y_T}\right)^\top$ are the estimated autocorrelation vectors of X_T and Y_T respectively, for some ℓ such that $\hat{\rho}_{i,X_T} \approx 0$ and $\hat{\rho}_{i,Y_T} \approx 0$ for $i > \ell$. The d_{ACF} is a typical way to investigate the data serial correlation being considered as the time domain approach. Another way is analyse the data dynamic in a frequency domain by using the distance based on the Periodogram function. The d_P distance is defined as [32]

$$d_P\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \frac{1}{n}\sqrt{\sum_{k=1}^{n} \left(I_{X_T}\left(\lambda_k\right) - I_{Y_T}\left(\lambda_k\right)\right)^2}, \tag{14}$$

where $I_{X_T}\left(\lambda_k\right) = T^{-1}\left|\sum_{t=1}^{T} X_t e^{-i\lambda_k t}\right|^2$ and $I_Y\left(\lambda_k\right) = T^{-1}\left|\sum_{t=1}^{T} Y_t e^{-i\lambda_k t}\right|^2$ are the periodograms of $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$, respectively, at frequencies $\lambda_k = 2\pi k/T, k = 1, \ldots, n$, with $n = [(T-1)/2]$. Defining $\phi(\cdot,\cdot) = d_{ACF}(\cdot,\cdot)$ or $\phi(\cdot,\cdot) = d_P(\cdot,\cdot)$, the V-based control rule in (11) will be called V_s.

3.3. V- based control charts

Let assume a group of n_1 reference batch samples available, i.e., time series $X_{i,t}$, for $i, \ldots, n_1$, of one process variable representing the in-control process. Each new batch sample Y_t is monitored simultaneously considering drifts and dynamics by using the V-based statistic. The V_d and V_s scores will be evaluated according to the limits set in the (11) using the false alarm α. For doing so, the step by step for setting the charts will be described as an algorithm as follows:

3.3.1. Steps for Drift monitoring

i) Compute the average of the n_1 time series, represented by $\overline{X}_t$, where $\overline{X}_t = (1/n_1)\sum_{i=1}^{n_1} X_{i,t}$.

ii) Define the mean centered time series data $X_{i,t}^c$ and Y_t^c, where $X_{i,t}^c = X_{i,t} - \overline{X}_t$, for $i = 1, \dots, n_1$ and $Y_t^c = Y_t - \overline{X}_t$ (smoothing serial correlation effects).

iii) Obtain the n_1 distances $d(\cdot,\cdot)$ in (12) for each pair (Y_t^c,$X_{i,t}^c$).

iv) Obtain the B_n and its variance by re-sampling procedure.

v) Obtain the V_d score by (11).

vi) Compare the V_d score with the quantile from the Standard Normal Distribution by using the false alarm probability of α.

3.3.2. Steps for Dynamics monitoring

i) Compute the moving average time series of length h from $X_{i,t}$, represented by $X_{i,t}^{ma}$, for $i = 1, \dots, n_1$ and Y_t^{ma} from Y_t.

ii) Define the moving average centered data $X_{i,t}^{cm}$ and Y_t^{cm}, where $X_{i,t}^{cm} = X_{i,t} - X_{i,t}^{ma}$, for $i = 1, \dots, n_1$ and $Y_t^{cm} = Y_t - Y_t^{ma}$ (smoothing trend effects).

iii) Obtain the n_1 distances $d_{ACF}(\cdot,\cdot)$ in (13) (or $d_P(\cdot,\cdot)$ in (14)) for each pair (Y_t^{cm}, $X_{i,t}^{cm}$) by using the *lag.max* values of the ACF (or Periodogram).

iv) Obtain the B_n and its variance by re-sampling procedure.

v) Obtain the V_s score by (11).

vi) Compare the V_s score with the quantile from the Standard Normal Distribution by using the false alarm probability of α.

4. Simulation study

4.1. Settings

In this Section, we generate batch processes in which the dynamic is described by the *Autoregressive* AR(1) and the *Moving Average* MA(1) models. In order to illustrate our method and compare with the MPCA-based approach, we present a Monte Carlo simulation presenting two kind of disturbances: (i) Disturbances in the serial correlation by using an AR(1) model (without intercept) representing the in-control trajectories of one process variable and disturbances following the θ parameter from MA(1) model (without intercept); (ii) Disturbances in the mean trajectory (drift changes) of the process representing by the ϕ_0 parameter from the AR(1) model with intercept. The AR(1) model with intercept and MA(1) model without intercept are respectively explicitly written as

$$x_t = \phi_0 + \phi_1 x_{t-1} + \epsilon_t \tag{15}$$

and

$$x_t = \epsilon_t + \theta_1 \epsilon_{t-1} \tag{16}$$

Table 1 shows the set of AR(1)/MA(1) parameters for in-control process and the simulation settings. New batches are generated considering scenarios with a wide range of disturbances in the intercept term ϕ_0 and the the θ_1 term from MA(1) model. We generate scenarios including numbers of batches with two time-length levels $T \in (20, 2000)$. Each scenario was replicated 1000 times. We do variate the number of reference batches I including three different levels, $I \in (50, 100, 1000)$. The control charts were setting to the false alarm probability of $\alpha = 0.05$. For each scenario 500 new batches were generated. The rate of batches beyond the control (r) and the Average Run Length index (ARL) were adopted to evaluate the chart's performance, where $ARL = 1/r$. The ARL_0 is the average number of batches until a false alarm (for $\alpha = 0.05$, $ARL_0 = 20$), i.e., points above control limits in the process without disturbances (in-control process). In contrast, ARL_1 is the average number of samples until an out-of-control batch falls outside the control limits. The former is a measure of the chart's sensibility. Simulations and calculations were conducted using R [33].

Table 1: Simulation settings

AR(1)/MA(1) coefficients	AR(1) settings (In-Control)	Disturbance levels	# Reference Batches	# New Batches	Batch length T	Run
ϕ_0 (AR)	1	$0, 0.5, 0.8, 1, 1.2, 1.5, 2$ (Simulation (ii))	50,100, 1000	500	20, 2000	1000
ϕ_1 (AR)	0.2					
θ_1 (MA)	-	$-0.5, -0.2, -0.1, 0.5$ (Simulation (i))				

Table 2 summarizes the results of the proposed charts and the MPCA-based charts for disturbances in the data serial correlation. The Table 2 shows the mean and standard deviation of ARL_0 values (highlighted in the gray line) for the in-control simulated batches [generated by the AR(1) without intercept with $\phi_1 = 0.2$] and ARL_1 for simulated batches with disturbance levels given by θ_1 from MA(1) model without intercept. In the first look, the $\mathcal{T}^2_{pca}$ and $\mathcal{Q}$ charts doesn´t fit to the ARL_0 nominal values when the number of time-instants are small but near or bigger than the number of batches. This is not new, since the MPCA-based is known to be suitable in scenario in which there are much more available batches than time-instants [10]. The $\mathcal{Q}$ pointed out so many false alarms with nearly all in-control batches beyond the limits, except for $I = 1000$ and $T = 20$, in this case the ARL_0 is close to the nominal one. The $\mathcal{T}^2_{pca}$ works near to the nominal false alarm only for $T = 20$, however, for $T = 2000$ we can noticed the poor sensibility of this chart. It is due the false alarm rate of points beyond the limits of $r = 0$ (ARL_0=1/0). For that reason, we can evaluate the ARL_1 performance of the $\mathcal{Q}$ chart for $T = 20$ and $I = 1000$ and the $\mathcal{T}^2_{pca}$ only for $T = 20$ and $I \in (50, 100, 1000)$. In the other hand, the proposed V_s and V_d charts are very close to the nominal target in all scenarios in terms of ARL_0, no matter if T are bigger than the I or conversely. Since we are disturbing the serial correlation, we are focused in the performance of the V_s chart. This chart outperforms the $\mathcal{T}^2_{pca}$ and $\mathcal{Q}$ charts in all comparable scenarios, except for the $T = 20$ and $I = 1000$ (which is the best scenario for those MPCA-based charts), for the disturbance levels of $\theta \in (-0.1, -0.2, -0.5)$, when the ARL_1 for the $\mathcal{Q}$ is a little smaller than the V_s. The V_d should be remain close to the nominal value, since there is no disturbance in the mean (drift). In a closer look, we see that the chart present ARL_1 close or bigger than the nominal value, or with missing ARL_1 values. Those big values and missing data means that the rate of points beyond the limits is $r = 0$ or very close to. This is not too bad, since there is no disturbance affecting drifts or trends, only the serial correlation.

Table 3 summarizes the results of the V_s and V_d and the MPCA-based charts for disturbances representing drifts in the mean of one process variable. The mean and standard deviation of ARL_0 values (highlighted in the gray line) for the in-control simulated batches [generated by the AR(1) with intercept $\phi_0 = 1$ and $\phi_1 = 0.2$] and ARL_1 for simulated batches generated by the same model, with disturbance levels in ϕ_0. The results are very close to the Table 2. Due the constrains of the MPCA-based charts for the time-instants small but near or bigger than the number of batches, we have here the same comparable scenarios as in the Table 2. Here, we are disturbing the mean, so the focus is in the performance of the V_d chart. As in Table 2, the proposed charts works well according to the nominal ARL_0 and the V_d outperform the benchmark for nearly all scenarios in terms of ARL_1, except for disturbances with $T = 20$ and $I = 1000$, again the best scenario for the MPCA-based charts. We can see that the V_s and V_d is in general very competitive in the MPCA-based favorable scenarios and work well when the number of time-instants is much bigger than the number of batches available in the reference data, as we can see in modern batch process.

Table 2: In-control process with AR(1) without intercept and $\phi_1 = 0.2$. Disturbances in the θ_1 coefficient of MA(1) model without intercept. Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1

I	T	θ_1	$V_s(\hat{\mu})$	$V_s(\hat{\sigma})$	$V_d(\hat{\mu})$	$V_d(\hat{\sigma})$	$\mathcal{T}^2_{pca}(\hat{\mu})$	$\mathcal{T}^2_{pca}(\hat{\sigma})$	$\mathcal{Q}(\hat{\mu})$	$\mathcal{Q}(\hat{\sigma})$
50	20	$\phi_1 = 0.2$	18.45	12.46	18.70	11.01	15.70	6.32	2.47	0.35
		0.5	13.03	6.14	41.49	28.30	14.37	7.67	2.10	0.29
		-0.1	4.15	1.34	318.71	181.09	63.28	54.03	1.36	0.08
		-0.2	3.45	1.44	286.34	187.39	44.97	33.18	1.26	0.06
		-0.5	2.12	0.58	45.76	56.48	18.08	10.17	1.10	0.03
	2000	$\phi_1 = 0.2$	24.43	13.38	25.62	20.63	–	–	1.00	0.00
		0.5	19.87	11.66	500.00	0.00	–	–	1.00	0.00
		-0.1	12.85	13.96	–	–	–	–	1.00	0.00
		-0.2	1.17	0.21	–	–	–	–	1.00	0.00
		-0.5	1.00	0.00	500.00	0.00	–	–	1.00	0.00
100	20	$\phi_1 = 0.2$	17.53	8.20	17.19	6.56	16.92	5.97	5.20	0.91
		0.5	14.18	7.29	33.83	19.44	11.03	2.97	4.41	1.10
		-0.1	4.00	0.95	349.03	153.03	37.34	19.30	1.68	0.12
		-0.2	3.22	0.70	291.26	151.56	26.90	13.70	1.47	0.09
		-0.5	1.96	0.32	40.31	29.93	10.18	3.03	1.18	0.04
	2000	$\phi_1 = 0.2$	18.43	6.29	22.79	9.72	–	–	1.00	0.00
		0.5	17.18	7.54	464.29	94.49	–	–	1.00	0.00
		-0.1	10.47	7.31	–	–	–	–	1.00	0.01
		-0.2	1.13	0.08	–	–	–	–	1.00	0.02
		-0.5	1.00	0.00	500.00	0.00	–	–	1.00	0.01
1000	20	$\phi_1 = 0.2$	15.09	2.97	15.73	3.10	20.32	4.11	28.13	7.56
		0.5	12.48	2.23	34.56	10.58	8.89	1.26	64.27	32.23
		-0.1	3.78	0.39	378.13	146.59	18.06	4.07	2.38	0.14
		-0.2	3.01	0.26	317.42	148.16	13.71	2.70	1.95	0.10
		-0.5	1.92	0.12	33.58	10.24	5.56	0.65	1.37	0.04
	2000	$\phi_1 = 0.2$	16.81	3.87	19.80	4.38	–	–	1.00	0.00
		0.5	16.04	3.13	458.33	102.06	–	–	1.00	0.00
		-0.1	7.53	1.60	–	–	–	–	1.00	0.00
		-0.2	1.10	0.03	–	–	–	–	1.00	0.00
		-0.5	1.00	0.00	500.00	0.00	–	–	1.00	0.00

Table 3: In-control process with AR(1) with intercept $\phi_0 = 1$ and $\phi_1 = 0.5$. Disturbances in the ϕ_0 coefficient. Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1

I	T	θ_1	$V_s(\hat{\mu})$	$V_s(\hat{\sigma})$	$V_d(\hat{\mu})$	$V_d(\hat{\sigma})$	$\mathcal{T}^2_{pca}(\hat{\mu})$	$\mathcal{T}^2_{pca}(\hat{\sigma})$	$\mathcal{Q}(\hat{\mu})$	$\mathcal{Q}(\hat{\sigma})$
50	20	0.0	22.74	29.64	1.41	0.16	1.84	0.40	1.86	0.19
		0.5	20.62	12.37	5.17	1.85	6.61	2.34	1.69	0.20
		0.8	17.52	10.48	13.57	5.90	14.06	4.79	1.48	0.18
		1.0	18.45	10.47	20.79	14.54	18.39	8.45	1.33	0.19
		1.2	18.45	11.96	14.15	6.25	14.21	5.26	1.27	0.18
		1.5	22.74	49.21	5.10	1.53	6.53	2.11	1.16	0.14
		2.0	18.50	11.02	1.45	0.19	1.97	0.45	1.06	0.09
	2000	0.0	21.55	9.64	1.00	0.00	–	–	1.00	0.00
		0.5	22.35	13.21	1.00	0.00	–	–	1.00	0.00
		0.8	19.46	9.08	3.47	0.94	–	–	1.00	0.00
		1.0	18.62	8.87	24.97	13.09	–	–	1.00	0.00
		1.2	20.55	10.26	3.74	1.19	–	–	1.00	0.00
		1.5	25.06	49.40	1.00	0.00	–	–	1.00	0.00
		2.0	21.77	12.98	1.00	0.00	–	–	1.00	0.00
100	20	0.0	18.40	16.42	1.39	0.09	1.71	0.25	3.72	0.49
		0.5	17.03	7.26	4.94	1.17	5.90	1.48	3.36	0.56
		0.8	17.36	11.38	13.16	4.83	12.64	3.85	2.82	0.63
		1.0	16.18	6.12	17.13	5.93	15.08	4.68	2.63	0.55
		1.2	17.05	8.47	12.92	4.72	12.31	3.47	2.39	0.61
		1.5	17.17	9.52	5.07	1.37	5.77	1.43	1.94	0.59
		2.0	16.85	7.02	1.40	0.12	1.74	0.28	1.53	0.47
	2000	0.0	18.91	6.70	1.00	0.00	–	–	1.00	0.00
		0.5	16.53	5.20	1.00	0.00	–	–	1.00	0.00
		0.8	18.76	7.29	3.40	0.63	–	–	1.00	0.00
		1.0	19.53	6.53	22.73	9.41	–	–	1.00	0.00
		1.2	20.95	11.34	3.35	0.63	–	–	1.00	0.00
		1.5	19.86	9.13	1.00	0.00	–	–	1.00	0.00
		2.0	18.77	6.73	1.00	0.00	–	–	1.00	0.00
1000	20	0.0	15.43	2.96	1.38	0.05	1.74	0.10	23.44	5.32
		0.5	15.40	3.58	4.68	0.52	7.06	0.98	21.35	4.47
		0.8	15.58	3.20	12.08	2.13	16.70	3.28	21.20	6.08
		1.0	15.72	3.87	16.33	2.83	21.09	4.60	19.73	4.35
		1.2	15.47	3.04	11.92	1.81	16.68	3.27	18.62	4.92
		1.5	15.17	2.99	4.80	0.54	7.36	1.05	16.87	5.81
		2.0	15.41	3.46	1.37	0.04	1.76	0.09	15.39	5.98
	2000	0.0	17.49	3.90	1.00	0.00	16.30	22.92	1.00	0.00
		0.5	16.45	3.23	1.00	0.00	–	–	1.00	0.00
		0.8	17.46	4.70	3.15	0.27	–	–	1.00	0.00
		1.0	17.15	3.63	20.58	5.48	–	–	1.00	0.00
		1.2	17.02	3.17	3.15	0.26	–	–	1.00	0.00
		1.5	16.76	3.23	1.00	0.00	–	–	1.00	0.00
		2.0	17.73	3.93	1.00	0.00	15.62	19.79	1.00	0.00

5. Application - Penicillium Fermentation

In order to illustrate the applicability of our methodology we consider data from penicillin fermentation batch process. This process generates batch data representing time series of process variables, such as temperatures, ph, dissolved oxygen, weight, etc. The data were downloaded from (http://www.industrialpenicillinsimulation.com/) The data set include three types of in-control batches available (including 30 batches each) and 10 batches presenting some faults resulting in process variables deviations. We choose one of these in-control batches, controlled by an Advanced Process Control (APC) solution using the Raman spectroscopy. The batches have uneven time length, so that the 835 time-instant was considered for all batches. Two process variables are chosen, one at a time (i.e., K=1), in order to illustrate our proposition. In the following subsections, the data are presented and the statistical analysis aimed to build the charts are described.

5.1. Oxigen Uptake Rate

Figures 1 and 2 show the time series data, the mean centered data, the moving average centered data and the ACF of the moving average centered data, respectively. The colors green and ruby represent the in-control group and the faulty group, respectively. Figure 3 shows the V_d and V_s control charts, respectively, set with the false alarm of $\alpha = 0.05$. The first 30 points represents the V_d (V_s) scores of the in-control batches and the 10 points after the horizontal dash line represents the V_d (V_s) scores of the monitoring faulty batches. The V_d chart pointing out six points above the limit. A closer look in the mean centered data on the right on Figure 1 can confirm the significant difference between the two groups. Some faulty batches (more specifically, six) present drifts mainly in the second half of the batch length. All faulty batches were detected by the V_s chart. Looking on the right on Figure 2, it understandable, since we can see clearly differences between the ACFs of two groups, manly in the lag 40, 50, 80, 150. The V_d (V_s) chart indicates three (two) false alarms, which is very close to the nominal false alarm probability of $\alpha = 0.05$.

5.2. Penicillin.concentration.P.g.L.

Figures 4 and 6 show the time series data, the same transformed data and the V_d (V_S) charts as the previous variable. The data on the right on Figure 4 shows a strong positive trend and we can notice that eight faulty batches are far different form the in-control green ones near to the end of the batches. The V_d chart (on the left of Figure 6) pointing out six points above the limit (among these eight). Two of them are not detected, since the time length of the uneven batches was restricted to $t = 835$. The V_s chart (on the left of Figure 6) signalized all the faulty batches. Again we can see differences between the two groups in the ACF chart on the right on Figure 5, manly in the lag 50, 80, 125. The two control charts are again with the false alarm rate close to the nominal value.

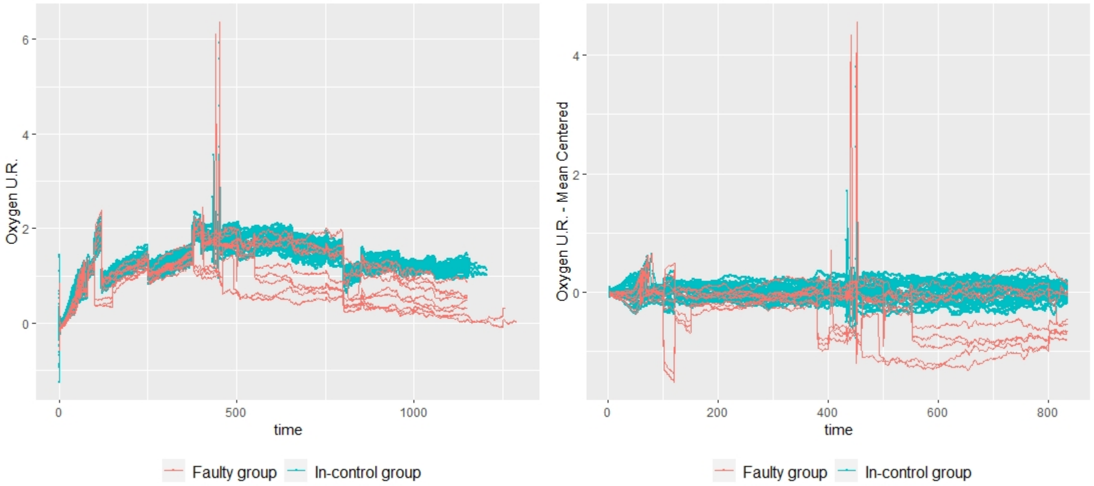

Figure 1: Oxygen Uptake variable: X_t, Y_t data and X_t^c, Y_t^c data

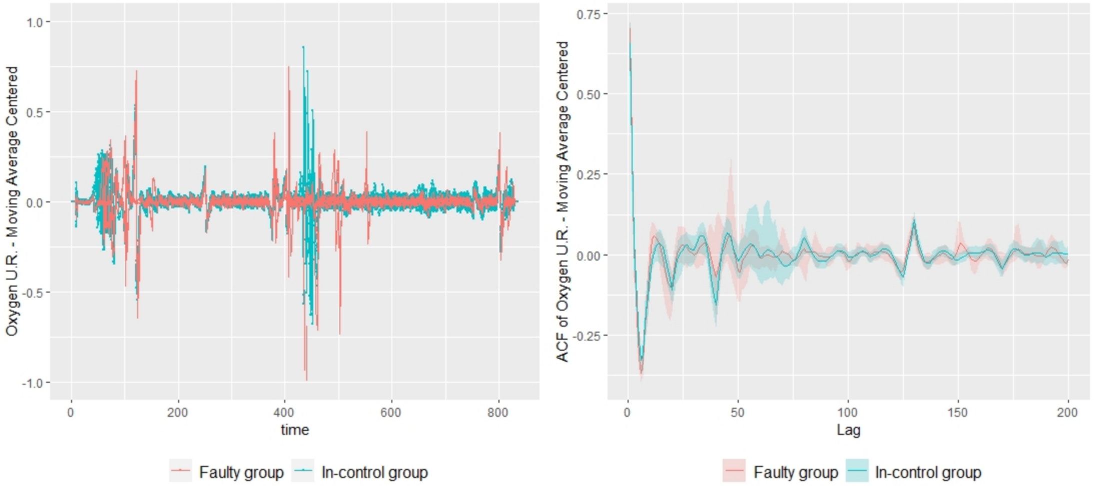

Figure 2: Oxygen Uptake variable: X_t^{cm}, Y_t^{cm} data and ACF of X_t^{cm}, Y_t^{cm} data

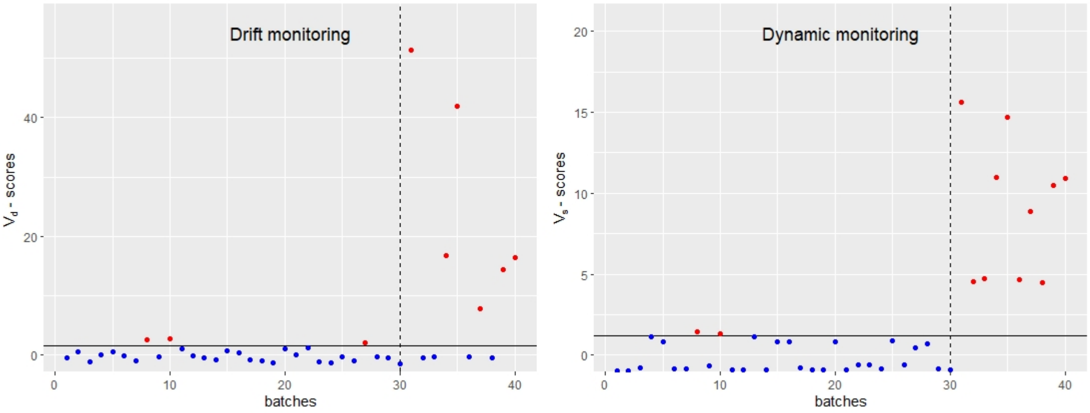

Figure 3: V_d and V_s control charts

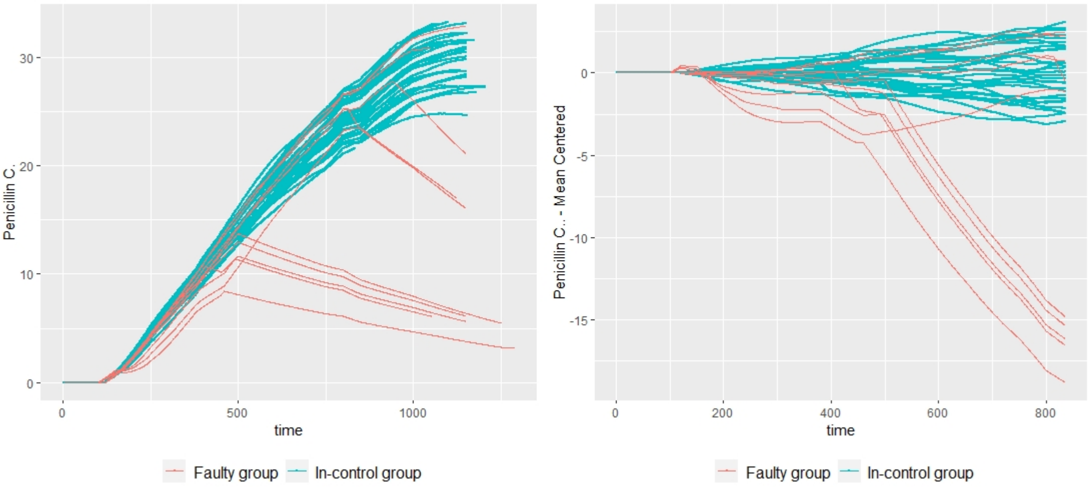

Figure 4: Penicillin Concentration variable: X_t, Y_t data and X_t^c, Y_t^c data

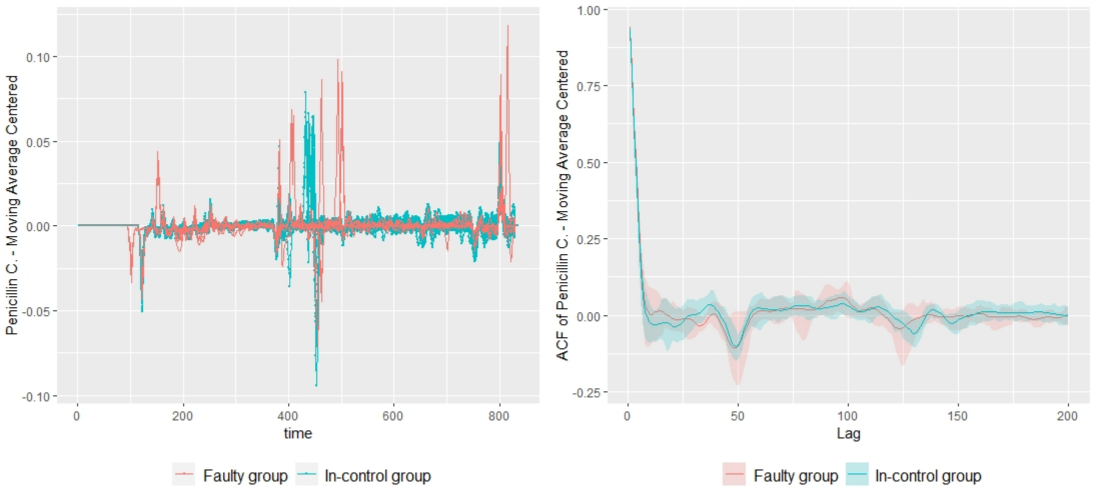

Figure 5: Penicillin Concentration variable: X_t^{cm}, Y_t^{cm} data and ACF of X_t^{cm}, Y_t^{cm} data

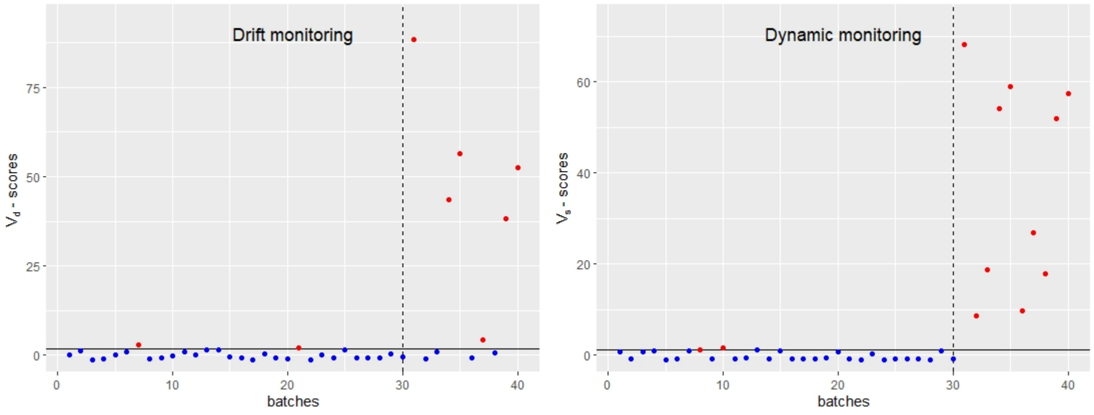

Figure 6: V_d and V_s control charts

6. Additional Issues

Some aspects involving the proposed methodology should be discussed to clarify its usefulness and applicability.

Through the simulation and the real data application we have shown the good performance of the U-based control charts for monitoring time series data considering serial correlation and drifts (including trends). For doing so, the euclidean distance between input vectors of the modified time-series data and the ACF distance (distance between input vectors with ACF values) are used. It's important to notice that under some suitable conditions (mentioned in the above paragraph), the U statistic remains asymptotically normal distributed in T and its properties are still valid for a variety of data functions, i. e., for different information in the input vectors. It means that we can generate V-charts to monitor a wide range of time series features through a variety of distance measures. Such a really flexible characteristic of that statistic opens its applicability.

The flexibility of the V-based charts allows us to make improvements to deal with batches generating multivariate time series, i. e., considering more than one variable under monitoring. In this context we can for instance develop the V-based chart to monitor the cross correlation between variables (two by two), using a distance measure in a vector of cross correlations values of lag g. The V_d chart for the drifts and the V_s chart for the individual serial correlation can also be extended to the multivariate case, using one chart for each variable separately or a chart based on one input vector of piled up values related to all variables jointly.

The reference batches available in phase $\mathcal{I}$ are considered to be the reflection of the in-control process in the proposed methodology. It means that those time-series represent the standard behavior of the variable in the process operating in a normal regime. However, the U-based theory presented allows us to develop a previous analysis in order to check if those reference batches are homogeneous in the statistical sense, i. e., we can run the U-based test to verify if all batches in the reference data set can be indeed considered as the reference group. It means that the data can not be significantly split in more than one distinct group. Otherwise, the significantly different batches could be removed from the initial group, in order to have a better characterization of the time-series variability in the reference group. This analysis can be useful in two situations: (a) We know that the batches coming from the in-control process (as in the simulation section and the real data application section) but we are intended to make sure that those batches are homogeneous in the statistical sense. It will increase the power of the U-based charts to detect differences in the new batch samples under monitoring; (b) We really don´t know if all of those available batches are generated from the process operating in the normal regime. So, in this case, we need to run the U-test to define one homogeneous reference group to apply the U-based control chart.

7. Conclusion

This paper proposed the new approach taking to account the batch-to-batch and the inner batch variability, based on the U-statistic theory. Trough the U-based statistics, named as V, we can compare time series data from new batches with the reference batches and evaluate then in terms of drifts (including trends) and serial correlation.

The V statistic is really flexible, since it doesn't depend on the input data distribution neither the time series model adjustments. Under week conditions, V is asymptotically Normal distributed in T, so we can build a suitable parametric control rule for modern batch process that generates large time-series data.

We've shown the good performance of the proposed V-based control chart, through the simulated and real batch data. We reinforce the flexibility of those charts to handle different batch data structures.

References

[1] P. Nomikos, J. F. MacGregor, Multivariate spc charts for monitoring batch processes, Technometrics 37 (1) (1995) 41–59.

[2] J. Chen, K.-C. Liu, On-line batch process monitoring using dynamic pca and dynamic pls models, Chemical Engineering Science 57 (1) (2002) 63–75.

[3] C. Undey, A. Cinar, Statistical monitoring of multistage, multiphase batch processes, IEEE Control systems magazine 22 (5) (2002) 40–52.

[4] T. Kourti, Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions, Journal of Chemometrics 17 (1) (2003) 93–109.

[5] J.-M. Lee, C. Yoo, I.-B. Lee, Fault detection of batch processes using multiway kernel principal component analysis, Computers & chemical engineering 28 (9) (2004) 1837–1847.

[6] Y. Yao, F. Gao, A survey on multistage/multiphase statistical modeling methods for batch processes, Annual Reviews in Control 33 (2) (2009) 172–183.

[7] L. Zhaomin, J. Qingchao, Y. Xuefeng, Batch process monitoring based on multisubspace multiway principal component analysis and time-series bayesian inference, Industrial & Engineering Chemistry Research 53 (15) (2014) 6457–6466.

[8] J. Peng, H. Liu, Y. Hu, J. Xi, H. Chen, Ascs online fault detection and isolation based on an improved mpca, Chinese Journal of Mechanical Engineering 27 (5) (2014) 1047–1056.

[9] J. Wang, W. Liu, K. Qiu, T. Yu, L. Zhao, Dynamic hypersphere based support vector data description for batch process monitoring, Chemometrics and Intelligent Laboratory Systems 172 (2018) 17–32.

[10] F. A. P. Peres, T. N. Peres, F. S. Fogliatto, M. J. Anzanello, Fault detection in batch processes through variable selection integrated to multiway principal component analysis, Journal of Process Control 80 (2019) 223–234.

[11] S. Wold, N. Kettaneh, H. Fridén, A. Holmberg, Modelling and diagnostics of batch processes and analogous kinetic experiments, Chemometrics and intelligent laboratory systems 44 (1-2) (1998) 331–340.

[12] J. Camacho, J. Picó, A. Ferrer, The best approaches in the on-line monitoring of batch processes based on pca: Does the modelling structure matter?, Analytica chimica acta 642 (1-2) (2009) 59–68.

[13] M. Jia, F. Chu, F. Wang, W. Wang, On-line batch process monitoring using batch dynamic kernel principal component analysis, Chemometrics and Intelligent Laboratory Systems 101 (2) (2010) 110–122.

[14] P. Van den Kerkhof, G. Gins, J. Vanlaer, J. F. Van Impe, Dynamic model-based fault diagnosis for (bio) chemical batch processes, Computers & chemical engineering 40 (2012) 12–21.

[15] C. Zhao, Concurrent phase partition and between-mode statistical analysis for multimode and multiphase batch process monitoring, AIChE Journal 60 (2) (2014) 559–573.

[16] C. Shang, F. Yang, X. Gao, X. Huang, J. A. Suykens, D. Huang, Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis, AIChE Journal 61 (11) (2015) 3666–3682.

[17] C. Shang, B. Huang, F. Yang, D. Huang, Slow feature analysis for monitoring and diagnosis of control performance, Journal of Process Control 39 (2016) 21–34.

[18] D. Marcondes Filho, L. P. L. de Oliveira, Multivariate quality control of batch processes using statis, The International Journal of Advanced Manufacturing Technology 82 (5-8) (2016) 867–875.

[19] Y. Qin, C. Zhao, S. Zhang, F. Gao, Multimode and multiphase batch processes understanding and monitoring based on between-mode similarity evaluation and multimode discriminative information analysis, Industrial & Engineering Chemistry Research 56 (34) (2017) 9679–9690.

[20] S. Zhang, C. Zhao, Slow-feature-analysis-based batch process monitoring with comprehensive interpretation of operation condition deviation and dynamic anomaly, IEEE Transactions on Industrial Electronics 66 (5) (2018) 3773–3783.

[21] X. Li, Z. Zhao, F. Liu, Latent variable iterative learning model predictive control for multivariable control of batch processes, Journal of Process Control 94 (2020) 1–11.

[22] Y. Wang, H. Yu, X. Li, Efficient iterative dynamic kernel principal component analysis monitoring method for the batch process with super-large-scale data sets, ACS omega 6 (15) (2021) 9989–9997.

[23] S. W. Choi, J. Morris, I.-B. Lee, Dynamic model-based batch process monitoring, Chemical Engineering Science 63 (3) (2008) 622–636.

[24] Y. Wang, J. Sun, T. Lou, L. Wang, Stability monitoring of batch processes with iterative learning control, Advances in Mathematical Physics 2017 (2017).

[25] D. Marcondes Filho, M. Valk, Dynamic var model-based control charts for batch process monitoring, European Journal of Operational Research 285 (1) (2020) 296–305.

[26] B. N. de Oliveira, M. Valk, D. Marcondes Filho, Fault detection and diagnosis of batch process dynamics using arma-based control charts, Journal of Process Control 111 (2022) 46–58.

[27] M. Valk, G. B. Cybis, U-statistical inference for hierarchical clustering, Journal of Computational and Graphical Statistics (2020).

[28] J. E. Jackson, A user's guide to principal components, Vol. 587, John Wiley & Sons, 2005.

[29] P. K. Sen, Robust statistical inference for high-dimensional data models with application to genomics, Austrian journal of statistics 35 (2&3) (2006) 197–214.

[30] A. Pinheiro, P. K. Sen, H. P. Pinheiro, Decomposability of high-dimensional diversity measures: Quasi-u-statistics, martingales and nonstandard asymptotics, Journal of Multivariate Analysis (2009).

[31] W. Hoeffding, A class of statistics with asymptotically normal distribution, The Annals of Mathematical Statistics (1948) 293–325.

[32] P. Montero, J. A. Vilar, Tsclust: An r package for time series clustering, Journal of Statistical Software 62 (2015) 1–43.

[33] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2020).
URL `https://www.R-project.org/`

7 Conclusão

Este trabalho apresentou uma nova abordagem para o monitoramento de processos em batelada, levando em conta tanto a variabilidade batelada a batelada quanto a variabilidade dentro de cada batelada, baseada na teoria de U-estatísticas. Por meio da abordagem proposta, compararam-se séries temporais de novas bateladas a um lote com bateladas sob controle, com o objetivo de identificar diferenças nos aspectos estruturais, como *drift* e correlação serial.

A principal vantagem desse método é sua flexibilidade, dado que não depende da distribuição dos dados ou de ajuste de modelos. Além disso, através de poucos pressupostos, a Carta de Controle V permite um monitoramento de diferentes estruturas de dados temporais isoladamente, sendo assim bastante adequada no contexto de processos em batelada.

Os resultados encontrados por meio de simulações, e a aplicação em dados reais, indicam um bom desempenho para a Carta de Controle proposta. De fato, a utilização da abordagem apresentada demonstrou ser eficiente sobretudo em contextos de *High Dimension Low Sample Size* (HDLSS), isto é, para poucas bateladas sob controle disponíveis e um número grande de medições das variáveis dentro da batelada.

Referências

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Sang Wook Choi, Julian Morris, and In-Beum Lee. Dynamic model-based batch process monitoring. *Chemical Engineering Science*, 63(3):622–636, 2008.

Gabriela B. Cybis, Marcio Valk, and Sílvia R. C. Lopes. Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, 2018.

Batista Nunes de Oliveira, Marcio Valk, and Danilo Marcondes Filho. Fault detection and diagnosis of batch process dynamics using arma-based control charts. *Journal of Process Control*, 111:46–58, 2022.

Manfred Denker. *Asymptotic distribution theory in nonparametric statistics.* Springer, 1985.

Donald Alexander Stuart Fraser. Nonparametric methods in statistics. 1956.

Zhiqiang Ge, Zhihuan Song, and Furong Gao. Review of recent research on data-based process monitoring. *Industrial & Engineering Chemistry Research*, 52 (10):3543–3562, 2013.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

Alan J Lee. U-statistics, volume 110 of statistics: Textbooks and monographs, 1990.

Erich Leo Lehmann. *Elements of large-sample theory.* Springer, 1999.

Danilo Marcondes Filho. Monitoramento de processos em bateladas através de cartas de controle multivariadas utilizando análise de componentes principais multidirecionais. 2001.

Danilo Marcondes Filho and Marcio Valk. Dynamic var model-based control charts for batch process monitoring. *European Journal of Operational Research*, 285(1):296–305, 2020.

Paul Nomikos and John F MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8):1361–1375, 1994.

Paul Nomikos and John F MacGregor. Multivariate spc charts for monitoring batch processes. *Technometrics*, 37(1):41–59, 1995.

Xia Pan and Jeffrey E Jarrett. Why and how to use vector autoregressive models for quality control: the guideline and procedures. *Quality & Quantity*, 46(3): 935–948, 2012.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 99:215–249, 2014.

Aluísio Pinheiro, Pranab Kumar Sen, and Hildete Prisco Pinheiro. Decomposability of high-dimensional diversity measures: Quasi-u-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, 2009.

Pranab Kumar Sen. Robust statistical inference for high-dimensional data models with application to genomics. *Austrian journal of statistics*, 35(2&3): 197–214, 2006.

Marcio Valk and Gabriela Bettella Cybis. U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*, 2020.

Marcio Valk and Aluísio Pinheiro. Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*, 33(4):608–619, 2012.

Erik Vanhatalo and Murat Kulahci. The effect of autocorrelation on the hotelling t2 control chart. *Quality and Reliability Engineering International*, 31 (8):1779–1796, 2015.

Yan Wang, Junwei Sun, Taishan Lou, and Lexiang Wang. Stability monitoring of batch processes with iterative learning control. *Advances in Mathematical Physics*, 2017, 2017.

Svante Wold, Nouna Kettaneh, Håkan Fridén, and Andrea Holmberg. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and intelligent laboratory systems*, 44(1-2):331–340, 1998.