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

Detecção e diagnóstico de falhas através de cartas de controle baseadas no modelo ARMA para monitoramento da dinâmica dos processos em bateladas

Batista Nunes de Oliveira

Porto Alegre, Setembro de 2021.

Dissertação submetida por Batista Nunes de Oliveira 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. Marcio Valk (PPGEst - UFRGS)

Coorientador:

Dr. Danilo Marcondes Filho (PPGEst - UFRGS)

Comissão Examinadora:

Dr. Fábio Mariano Bayer (UFSM)

Dr. Michel José Anzanello (UFRGS)

Dr. Ângelo Marcio Oliveira Sant'anna (UFBA)

Data de Apresentação: 28 de Setembro de 2021

Agradecimentos

Agradeço:

- Aos meus pais Maria Geny e Oscar (em memória), pelos ensinamentos primordiais e caráter, além de incentivo nos meus estudos.
- A minha esposa Margarete que tantas vezes não pude dar a devida atenção.
- Ao meu filho Renan pelas conversas e diálogos a respeito das possibilidades estatísticas.
- Ao prof. Dr. Marcio Valk, pela orientação, dedicação, paciência e por ter compartilhado tanto conhecimento.
- Ao prof. Dr. Danilo Marcondes Filho, meu coorientador, por todas as dicas e empenho pra que esse trabalho se realizasse.
- Aos membros da banca, pela disponibilidade.
- Aos colegas, funcionários e professores do Programa de Pós-Graduação em Estatística da UFRGS, pela convivência e amizades feitas.
- A Deus que está conosco e ilumina nossos caminhos.

Resumo

Processos em bateladas são amplamente usados na produção de uma grande variedade de itens, principalemente em indústrias químicas, bioquímicas, farmacêuticas e alimentos. Este tipo de processo disponibiliza em cada batelada amostras de séries temporais que descrevem sucessivas medições de variáveis que sinalizam sobre o andamento deste. Abordagens tradicionais de CEP(controle estatístico de processo) aplicadas a estes processos não utilizam modelos de séries temporais, pois a teoria inferencial de tais modelos não está construída baseada em replicações de séries temporais. Para monitoramento da dinâmica das variáveis de tais processos, usando como referência séries temporais disponíves, propomos um conjunto de cartas de controle baseadas nas estatísticas de *Hotelling* e *t-Stutent* modificadas, as quais acomodam as estimativas obtidas pelo ajuste do modelo ARMA. Esse controle se dá diretamente nos coeficientes do ARMA, que se diferência das abordagens clássicas. Adicionalmente, esta abordagem fornece informações que auxiliam no diagnóstico das alterações na dinâmica detectadas pelas cartas de controle. Implementamos simulações e uma situação com dados reais, nas quais os resultados constatam o bom desempenho da nossa abordagem.

Índice

1 Introdução

Processos em bateladas são amplamente usados na produção de uma grande variedade de itens. A estrutura de dados destes processos contém amostras de séries temporais representando sucessivas medições de um conjunto de variáveis de análise. Os dados oriundos destes processos tem características dinâmicas marcantes. Abordagens tradicionais não levam em conta diretamente a característica temporal dos dados e se baseiam no artigo precursor de [Nomikos and MacGregor, 1995], decompondo a estrutura tridirecional (I-bateladas $\times$ K-variáveis $\times$ T-instantes de tempo) em um arranjo bidirecional (I-bateladas $\times$ KT-variáveis/instantes de tempo). Em linhas gerais, tomando um conjunto de bateladas históricas como repetições, utilizam técnicas multivariadas no domínio variável-tempo para modelar tais dados e construir cartas de controle. Estas técnicas aplicadas neste arranjo de dados capturam em algum nível as correlações seriais nas variáveis e a característica dinâmica de tais dados.

A literatura de Controle Estatístico de Processo (CEP) é restrita em relação à proposição de cartas de controle para monitoramento de processos em bateladas utilizando modelos de séries temporais. A principal razão se deve ao fato de que a teoria inferencial no contexto de séries não está baseada em replicação de séries temporais. Particularmente, a estimação dos parâmetros do modelo bem como a distribuição amostral dos estimadores é derivada a partir da replicação de observações ao longo de uma única série. Em resumo, a variabilidade no domínio do tempo é modelada. Dessa forma, precisamos adaptar tal teoria para incorporar a variabilidade entre bateladas à teoria inferencial clássica. Em outras palavras, os estimadores dos parâmetros e suas distribuições devem incorporar a variabilidade entre amostras de séries temporais disponíveis. Existe um número restrito de trabalhos que tratam desse problema. No artigo da dissertação destacamos alguns que trazem contribuições propondo abordagens baseadas em modelos de séries temporais que buscam combinar a variabilidade dentro da batelada (análise usual no domínio do tempo) e a variabilidade entre bateladas.

Esta dissertação apresenta uma abordagem baseada no modelo autorregressivo de médias móveis (ARMA) [Box et al., 1970], para monitorar a dinâmica dos processos em bateladas. Um grupo de cartas de controle para monitorar os parâmetros do modelo através das estimativas obtidas de bateladas históricas é apresentado. As cartas estão baseadas em modificações propostas das distribuições de *Hotelling* e *t-Stutent* para acomodar o conjunto de estimativas dos parâmetros geradas das bateladas. Adicionalmente, a abordagem proposta permite diagnosticar os descontroles sinalizados nas cartas, fornecendo indicativos sobre quais coeficientes estimados foram mais afetados pelos distúrbios. Desenvolvimentos teóricos, estudos de simulação e uma aplicação são apresentados neste trabalho. Os resultados destas sinalizam o bom desempenho da abordagem.

Este trabalho está organizado da seguinte forma: Nesta Seção 1 apresentamos uma introdução contextualizando o tema; na Seção 2 descrevemos os objetivos do presente trabalho; uma descrição sobre o CEP no contexto de processos em bateladas é apresentada na Seção 3; o modelo ARMA é brevemente apresentado na Seção 4; a Seção 5 descreve a *Carta de Controle de Hotteling*; a Seção 6 resumimos a abordagem proposta; na Seção 7 apresenta o artigo da dissertação e por fim na Seção 8 a conclusão.

2 Objetivo

2.1 Objetivo Geral

Desenvolver um conjunto de cartas de controle baseadas no modelo ARMA para monitorar a dinâmica de processos em bateladas.

2.2 Objetivos Específicos

- Apresentar uma breve revisão de literatura sobre abordagens para monitoramento de processos em bateladas baseadas em modelos de séries temporais.
- Desenvolver uma abordagem de controle baseada no modelo ARMA para monitoramento de processos em bateladas, baseado na estatística de $Hotelling$ modificada aplicada aos coeficientes do modelo ARMA.
- Estender a abordagem permitindo diagnóstico, buscando identificar quais coeficientes do ARMA possivelmente são a origem do descontrole, através da estatística $t - Student$ modificada.
- Realizar estudo de simulação comparando o desempenho da abordagem proposta com abordagens baseadas nos resíduos do modelo ARMA.

3 CEP Tradicional para Processos em Bateladas

Processos em batelada são utilizados em diversos setores industriais (por exemplo, na manufatura de alimentos e fármacos). Nesses processos, matérias-primas são carregadas em uma unidade de processamento e submetidas a uma série de transformações até a obtenção do produto final. O desempenho do processo é descrito por variáveis, monitoradas ao longo da batelada. Dados resultantes desses processos tendem a apresentar uma característica dinâmica descrita pela estrutura de correlação serial e cruzada das variáveis do processo.

Os processos em bateladas típicos apresentam as seguintes macro etapas:

(i) Uma certa combinação de produtos ou processos de manufatura são realizados.

(ii) Durante a batelada, os produtos sofrem uma série de transformações e sensores medem com frequência uma série de variáveis do processo, tais como temperatura, pressão e taxas de mistura.

(iii) Após a batelada terminada, o produto final é analisado para garantia das características de qualidade.

Os dados dos processos em bateladas possuem uma estrutura tridirecional ($I \times K \times T$), onde I indica o número de bateladas, K o número de variáveis e T o número de instantes de tempo. Dessa forma, em cada batelada temos uma série temporal K-dimensional de tamanho T disponível, contendo dados com uma forte característica dinâmica.

As abordagens tradicionais para monitoramento de processos em bateladas não estão baseadas em modelos de séries temporais. Estas buscam capturar a característica dinâmica dos dados utilizando técnicas de estatística multivariada em arranjos bidirecionais dos dados. A técnica mais amplamente usada é a Análise de Componentes Principais (ACP) ([Hotelling, 1933];[Pearson, 1901]). Escores da ACP são obtidos e a Carta de Controle baseada na distribuição de *Hotteling* [Hotelling, 1947] é construída para monitorar as médias dos escores. O desdobramento mais comum é descrito na Figura 1 (a). Este arranjo proposto no trabalho precursor de [Nomikos and MacGregor, 1995] assume uma estrutura bidirecional do tipo ($I \times KT$). Utilizando bateladas como repetições (linhas da matriz de dados), a ACP é aplicada nas KT colunas da matriz. A variabilidade entre bateladas é modelada dessa forma e a característica dinâmica dos dados é capturada em algum nível. Note que neste arranjo as correlações seriais e cruzadas estão confundidas, deste modo os descontroles sinalizados nas cartas de controle tornam-se difíceis de determinar na etapa de diagnóstico. A Figura 1 (b) apresenta uma arranjo alternativo proposto no trabalho de [Wold et al., 1998]. Este arranjo estrutura os dados numa matriz ($IT \times K$). A ACP neste caso é aplicada nas K variáveis originais considerando IT linhas da matriz como repetições. Esta abordagem prioriza as correlações médias entre variáveis no conjunto dos instantes $\times$ bateladas. Perde-se a característica dinâmica e o foco é identificar variáveis mais influentes na variabilidade geral dos dados (isto é, foco no diagnóstico dos descontroles). [Camacho et al., 2009] apresentam um estudo comparativo destas abordagens apontando pontos positivos e negativos de ambas.

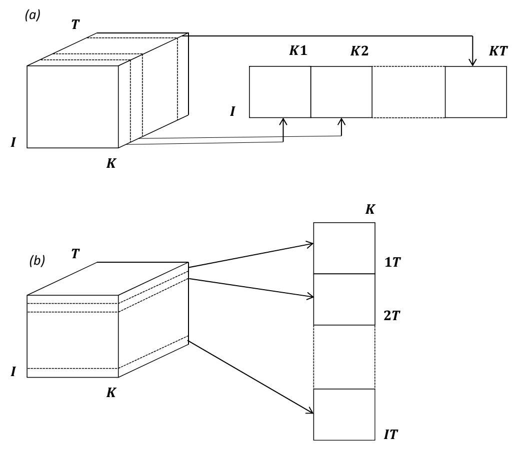

Figura 1: (a) Arranjo no domínio do tempo: $I \times KT$ proposto por Nomikos e Macgregor (1995). (b) Arranjo no domínio das variáveis : $IT \times K$ proposto por Wold et al. (1998).

4 Modelos ARMA

Os modelos ARMA são uteis para descrever séries temporais advindas de processos estocásticos fracamente estacionários. Para uma leitura sobre o assunto, recomendamos o livro de Brockwell and Davis (1990). Os modelos ARMA tem como principais características a versatilidade ao não necessitar de covariáveis para explicar a variável dependente, possibilitar a predição e também previsão de valores futuros. Além disso, as propriedades estatísticas desses modelos são amplamente exploradas e popularizadas na literatura e sua utilização muito comum nas áreas da economia, nas ciências físicas e geofísicas, sendo essas interessadas tanto no espectro do processo, quanto nos estimadores e na previsão. Os modelos ARMA combinam a característica do autorregressivo (AR), o qual envolve regredir a variável nos seus próprios valores defasados, com a do média móvel (MA) que considera uma combinação linear de defasagens do termo de erro como regressores do modelo. De forma geral, o modelo ARMA(v,w) pode ser escrito por

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_v x_{t-v} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_w \epsilon_{t-w}, \quad t \in Z, \tag{1}$$

ou ainda, na forma compacta

$$\Phi(L)x_t = \Theta(L)\epsilon_t, \tag{2}$$

em que ϵ_t é um processo ruído branco com média zero e variância σ^2, denotado por RB(0,σ^2) e L é o operador defasagem com a seguinte característica, $L^v x_t = x_{t-v}$. $\Phi(\dot{})$ e $\Theta(\dot{})$ são os polinômios característicos autorregressivo e de médias móveis, respectivamente, dados por

$$\begin{aligned} \Phi(z) &= 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_v z^v, \ \ \phi_v \neq 0; \\ \Theta(z) &= 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_w z^w, \ \ \theta_w \neq 0, \end{aligned}$$

em que z é um número complexo.

Duas propriedades dos polinômios característicos são importantes para que os métodos de estimação dos modelos ARMA tenham propriedades estatísticas confiáveis. O primeiro conceito é o de causalidade, o qual é garantido se, e somente se as raízes de $\Phi(z)$ estão fora do círculo unitário, ou seja,

$$|z| \leq 1 \rightarrow \Phi(z) = 1 - \phi_1 z - \cdots - \phi_v z^v \neq 0.$$

Outro conceito importante é o de invertibilidade, o que ocorre se, e somente se, as raízes de $\Theta(z)$ estão fora do círculo unitário, isto é,

$$|z| \leq 1 \rightarrow \Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_w z^w \neq 0.$$

4.1 Ajuste do Modelo e Propriedades dos Estimadores

As funções de autocorrelação (ACF) e de autocorrelação parcial (PACF) podem ser utilizadas para determinar as ordens v e w do modelo. As características de decaimento dos valores amostrais da ACF e da PACF e de quais lags são significativamente diferente de zero indicam qual modelo é mais adequado. A utilização do critério de informação Akaike (AIC) [Akaike, 1974] e do critério bayesiano (BIC) [Schwarz, 1978] é também recomendado para determinar o melhor modelo, juntamente com a análise de resíduos.

Uma vez escolhidos v e w, restam as estimações dos parâmetros do modelo, i.e., a média dos coeficientes $\{\phi_i, \theta_j : i = 0, \ldots, v, \ j = 1, \ldots, w\}$ dado as I bateladas e a variância do ruído branco σ^2. É usual assumir que a média da série temporal a ser modelada seja zero (além de ser fracamente estacionária) e caso não tenha média zero, modela-se a série temporal ajustada pela média $\{x_t - \overline{x}\}$.

Para uma série temporal advinda de um processo estocástico causal e invertível, em que os polinômios $\Phi(\cdot)$ e $\Theta(\cdot)$ não possuem zeros em comum e $\epsilon_t \sim IID(0, \sigma^2)$, o estimador de mínimos quadrados de $\hat{\boldsymbol{\beta}} = (\phi_1, \ldots, \phi_v, \theta_1, \ldots, \theta_v)$ satisfaz

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \dot{\sim} N(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})),$$

em que $\dot{\sim}$ denota distribuição assintótica e $\boldsymbol{V}(\boldsymbol{\beta})$ é a matriz de covariância assintótica (ver Teorema 8.11.1 de Brockwell and Davis (1990)).

5 Carta de Controle de Hotteling

As primeiras publicações na perspectiva multivariada foram feitas por [Hotelling, 1947], utilizando abordagem multivariada em dados contendo informações sobre bombardeios durante a Segunda Guerra Mundial. Seu uso foi difundido, pois observou-se que a maioria das variáveis de monitoramento do processo não se comportam de forma independente, portanto não sendo adequado tratá-las utilizando cartas de controle univariadas. A *carta de controle de Hotteling* é a extensão multivariada da carta de controle de Shewhart e permite o monitoramento conjunto de um vetor de médias.

Considere $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ um vetor aleatório com distribuição normal K-variada. Num processo sob controle assumimos o vetor de médias $\boldsymbol{\mu} = \boldsymbol{\mu_0}$ e a matriz de covariâncias $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_0}$. Na prática, na fase $\mathcal{I}$ da elaboração da carta de controle, $\boldsymbol{\mu_0}$ e $\boldsymbol{\Sigma_0}$ são estimados a partir de amostras aleatórias independentes obtidas de $\boldsymbol{X}$. A carta de controle clássica utilizada para monitoramento do vetor de médias é baseada na estatística $\mathcal{T}^2$ de *Hotteling* [Montgomery, 2007]. Na fase $\mathcal{II}$, considerando novas amostras aleatórias univariadas do vetor $\boldsymbol{X}$, $\mathcal{T}^2$ apresenta a seguinte distribuição:

$$\mathcal{T}^2 = (\mathbf{X} - \overline{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X} - \overline{\mathbf{X}}) \sim \frac{K(I-1)}{(I-K)} F_{K, I-K} \tag{3}$$

$\mathcal{T}^2$ segue uma distribuição F de Snedcor e $\overline{\mathbf{X}}$ e $\mathbf{S}$ representam respectivamente o vetor de médias e a matriz de covariâncias considerando o número I de amostras históricas do processo sob controle. A carta de controle $\mathcal{T}^2$ é construída e o limite de controle é determinado considerando o percentil da distribuição F a partir da probabilidade de alarme falso $\alpha = 0.01$. Em cada nova amostra, o escore $\mathcal{T}^2$ é obtido em (3). Escores acima do limite de controle sinalizam que a amostra foi obtida do processo onde $\boldsymbol{\mu} \neq \boldsymbol{\mu_0}$. Em resumo, há evidências de que o vetor de observações foi obtido de um processo no qual as variáveis (ou um subgrupo $< K$) possuem valores significativamente diferentes dos seus valores esperados.

A estatística escrita na expressão (3) é utilizada como base para a carta de controle $\mathcal{T}^2$ de Hotelling. Na fase $\mathcal{II}$, os limites de controle para a estatística $\mathcal{T}^2$ são os seguintes:

$$\begin{aligned} LSC &= F_{(K,I-K),\alpha}; \\ LIC &= 0, \end{aligned}$$

onde I é o número de amostras preliminares obtidas no processo sob controle e α representa o percentil da distribuição de Snedecor.

6 Resumo da Abordagem Proposta

Neste trabalho iremos apresentar uma abordagem para monitoramento de processos em bateladas através da proposição de cartas de controle baseada no modelo ARMA de séries temporais.

A proposta busca adaptar a tradicional carta de controle de Hotteling para monitorar coeficientes estimados do ajuste do modelo ARMA aos dados das bateladas. Para tanto propomos uma modificação das estatísticas de *Hotelling* e *t-Stutent* para acomodar o conjuto de estimativas dos coeficientes gerados em cada batelada sob controle disponível na fase de construção das cartas de controle.

Construímos então a carta de controle $\mathcal{T}^2$ para monitorar os coeficientes ajustados de bateladas novas. Adicionalmente, propomos a carta t_β para fins de diagnóstico de alterações sinalizadas pela carta $\mathcal{T}^2$. A base teórica de tal abordagem é detalhada no artigo (seção 7), incluindo estudos de casos envolvendo dados simulados e dados reais.

7 Artigo

Autores: Batista Nunes de Oliveira, Marcio Valk, Danilo Marcondes Filho

Título: Fault detection and diagnosis of batch process using dynamic ARMA-based control charts

Ano: 2021

Fault detection and diagnosis of batch process using dynamic ARMA-based control charts

Batista Nunes de Oliveira, Marcio Valk, Danilo Marcondes Filho*

Department of Statistics, Federal University of Rio Grande do Sul, Av. Bento Gonçalves, 9500 – Porto Alegre, RS, Brazil

Abstract

A wide range of approaches for batch processes monitoring can be found in the literature. This kind of process generates a very peculiar data structure, in which successive measurements of many process variables in each batch run are available. Traditional approaches do not take into account the time series nature of the data. The main reason is that the time series inference theory is not based on replications of time series, as it is in batch process data. It is based on the variability in a time domain. This fact demands some adaptations of this theory in order to accommodate the model coefficient estimates, considering jointly the batch to batch samples variability (batch domain) and the serial correlation in each batch (time domain). In order to address this issue, this paper proposes a new approach grounded in a group of control charts based on the classical ARMA model for monitoring and diagnostic of batch processes dynamics. The model coefficients are estimated (through the ordinary least square method) for each historical time series sample batch and modified *Hotelling* and *t-Student* distributions are derived and used to accommodate those estimates. A group of control charts based on that distributions are proposed for monitoring the new batches. Additionally, those groups of charts help to fault diagnosis, identifying the source of disturbances. Through simulated and real data we show that this approach seems to work well for both purposes.

Keywords: Batch processes monitoring, ARMA model, ARMA control charts, Modified *Hotelling* and *t-Student* distribution

1. Introduction

Industrial batch processes are used to produce a wide range of products. This kind of process generates for each batch run time series representing multiple measurements of process variables. That peculiar data

*Corresponding author
Email address: `marcondes.filho@ufrgs.br` (Danilo Marcondes Filho)

structure containing samples of time series and data with a strong dynamic feature makes it still very challenging to develop control charts based on time series models.

Traditional monitoring approaches do not take into account directly the time series nature of the data. Most of them decompose the tri-dimensional data array (batches $\times$ variables $\times$ time-instants) in a two-dimensional array (batches $\times$ variables/time-instants), based on the precursor approach of [1]. In this context, considering batches as sample replications, control approaches are proposed by using multivariate techniques (as Principal Components, Partial Least Squares, Discriminant Analysis, Support Vector machines, Neural Networks, etc) applied in the variable/time domain. These multivariate-based control charts are able to capture the dynamic data behavior in some way. We can mention a number of papers presenting improvements, applications and fruitful discussions in this direction in [2], [3], [4], [5], [6], [7], [8], [9], [10] and [11].

Control charts based on time series models are well known in the context of continuous processes. These kinds of processes have an intrinsic two-way data structure (samples $\times$ variables) since there are no replications of measures in each sample, i.e, there is no time dimension. Those models are mainly aimed to deal with the serial sample correlation, by using traditional control charts for the uncorrelated residuals or the cumulative-based charts for the fitted values. In both cases, as the first step, the model is adjusted from historical in-control samples and, in the next step, future samples are monitored through those charts. To accomplish that goal there are a wide range of propositions using the Autoregressive Moving Average (ARMA) and Vector Autoregressive (VAR) models. The good illustration can be seen in a case study presented in [Montgomery, 2007], in which a Shewhart control chart is used to monitor the residuals of the AR(1) model. In the context of a multivariate continuous process, the same goal is accomplished by using the VAR model for uncorrelated data and the Hotelling-based control chart for monitoring the vector of residuals. Some papers presenting the ground theory and additional contributions in such direction can be found in [12], [13], [14], [15] and [16]. We can also find in the literature approaches for monitoring fitted values from those models. In this case, fitted values from new samples are monitored by using the *exponentially weighted moving average* (EWMA) based control charts. A good review of those procedures can be seen in [17], [18], [19], [20], [21], [22], [23], [24], [25] and [26].

There are a few works in the literature presenting approaches for batch process monitoring based on time series models. The reason why this topic indeed hasn't been fully explored yet is that the time series inference theory is not based on replications of time series (each one bringing successive measurements of processes variables), as it is in batch process data. It is based on the variability in a time domain. This fact demands some adaptations of this theory in order to accommodate the coefficient estimates, taking into account the number of batch samples (batch to batch variability - batch domain analysis). The main problem is to build a single estimate for the model coefficients given a number of time series available, combining information from the batch and time domains. We highlight the work of [27], which propose a set of charts based on the traditional Hotelling statistic for the VAR residuals and fitted vector of observations, obtained thought the adjusted VAR from the historical time series samples batches in a reduced variable space. In this approach, the VAR coefficient estimates are done by using the Partial Least Square regression technique instead of by using the VAR estimation theory. [28] propose a group of control charts based on the 2D ARMA model. That model formulation try to capture the within-batch and batch-to-batch variability. For each batch they use the *iterative step-wise regressions* (SWR) and the *least absolute shrinkage and selection operator* (LASSO) to identify the model order and select the model coefficients. Future batch samples are monitored through control charts based in two stability index built from the coefficients estimates (the batch-to-batch and the within-batch index). [29] presented a recent approach to deal with batch processes using VAR models focused on the VAR coefficients directly. In short, the VAR coefficients are estimated for each historical time series sample batch and by using a single estimate [as a combination of individual ordinary least square (OLS) estimates from each batch], the *Hotelling* and the *Generalized Variance* control charts are used for monitoring new batches.

This paper proposes a new approach grounded in a group of control charts based on the classical ARMA model for dynamic monitoring and diagnostic of batch processes. Considering one variable at the time, we present a control approach based on the ARMA coefficients. The model coefficients are estimated for each historical time series sample batch and the control charts based on the modified *Hotelling* distribution are used to monitor new batches. Additionally, we extend this idea by using a group of control charts, one for each ARMA coefficient, based on the modified *t-Stutent* distribution, in order to help the diagnostic of disturbances detected by the *Hotelling* chart. There are two meaningful contributions in our proposition. The first one is that through the modified *Hotelling* and *t-Stutent* distributions the model coefficient estimates generated from the number of historical batches can be easily accommodated. There is no need to use any estimation method based on complex algorithms and so deal with convergence and computational time issues. Also, the derived exact distributions makes the control charts capable to detect disturbances of any level, even in a scenario in which there are few in-control batch samples available. The second one is that, unlike these mentioned approaches, we address the fault diagnose problem by using a group of *t-Stutent* control charts. We show through a simulated batch process that the proposed approach outperforms a competitor based on the model residuals (the most common approach used in the continuous and batch processes) and it is powerful in terms of disturbance diagnosis. Furthermore, this approach seems to work well when applied in a real data set.

In order to describe our proposition, the paper is organized as follows: Section 2 brings a detailed description of our methodology, including the basis of ARMA models, the *Hotelling* and *t-Stutent* modified statistics and the ARMA-based control charts. In Section 3, the proposed approach is illustrated through simulated batch data. Section 4 shows an application in a real data set. Conclusions are presented in Section 5.

2. ARMA-based control approach

2.1. ARMA model

The Autoregressive-Moving-Average (ARMA) model is widely used in time series analysis and forecasting due to the flexibility and suitable statistical properties. The class of ARMA model was popularized by [30] and is characterized by a simple and parsimonious formulation. It combines the Autoregressive (AR) model, which involves regressing a variable on its own lagged values, with moving averages (MA) model, which considers the error term as a linear combination of its own lagged terms. In general, we can write the ARMA model as follows:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_v x_{t-v} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_w \epsilon_{t-w}, \quad t \in \mathbb{Z}, \tag{1}$$

where the error term is a white noise process (WN) with zero mean and variance σ^2, noted by $\epsilon_t \sim WN(0, \sigma^2)$, $t \in \mathbb{Z}$. This formulation is usually referred as ARMA(v, w) since v lags of the return are used as well as w lags of the error term to specify the linear functional form to be estimated. Let's consider the vector of parameters

$$\boldsymbol{\beta} = [\phi_0, \phi_1, \ldots, \phi_v, \theta_1, \ldots, \theta_w]. \tag{2}$$

Assuming that the process $\{x_t\}$ defined in (1) is causal and invertible, than for time-series of length T sampled from this process, the asymptotic distribution (in T) of the OLS estimators $\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = [\widehat{\phi}_0, \widehat{\phi}_1, \ldots, \widehat{\phi}_v, \widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_w]$, is given by Theorem 8.11.1 in [31]. Under suitable conditions, it follows that

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \dot{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{3}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ plays the role of the variance and covariance matrix of the $\boldsymbol{\beta}$ estimators and $\dot{\sim}$ means asymptotic convergence in distribution, when T increases.

Unfolding (3) we can write the univariate asymptotic distribution of each individual element of the vector $\widehat{\boldsymbol{\beta}}$ in (2). Let $\widehat{\beta}_*$ be any element of vector $\widehat{\boldsymbol{\beta}}$, than we have

$$(\widehat{\beta}_* - \beta_*) \dot{\sim} N_p(0, \sigma^2_{\beta_*}), \tag{4}$$

where $\sigma^2_{\beta_*}$ is the corresponding element of the main diagonal of the $\mathbf{\Sigma}_{\boldsymbol{\beta}}$.

In case of data coming from normally distributed variable, we consider the exact distribution of OLS estimated coefficients in (3) and (4) rather than asymptotic one.

2.2. Hottelling and t-Student adjusted distributions

Following the aim of our work, we now assume the scenario with many trials available from an ARMA process, i.e., data samples representing time series. This is a typical context of a batch process that generates data representing trajectories of a variable to be considered under monitoring. In the next Section, the set of ARMA-based control charts for this kind of process will be proposed. They are based on the adaptation of the classical *Hotelling* $\mathcal{T}^2$ statistic [32]. In the two theorems below we demonstrate the distribution of the quantities that are the ground of our approach.

Theorem 1. *Let's consider I time series of length T from an ARMA(v,w) process in* (1)*, $\boldsymbol{\beta}$ the vector of parameters defined in* (2) *and $\widehat{\boldsymbol{\beta}}_i$ the vector of OLS estimates for the i^{st} sample, both vectors of length $p = v + w + 1$. Assume that $\widehat{\beta}_{*i}$ is an element $\widehat{\boldsymbol{\beta}}_i$. Than,*

$$\frac{\left(\widehat{\boldsymbol{\beta}}_{*i} - \widehat{\overline{\boldsymbol{\beta}}}_*\right)}{S_{\widehat{\beta}_{*i}}} \dot{\sim} \sqrt{\frac{(I+1)}{I}} t_{(I-1)}, \ \ \textit{as } T \ \textit{ increases}, \tag{5}$$

where $S^2_{\widehat{\beta}_{*i}} = \frac{1}{(I-1)} \sum_{i=1}^{I} \left(\widehat{\beta}_{*i} - \widehat{\overline{\beta}}_*\right)^2$, $\widehat{\overline{\beta}}_* = \frac{1}{I}\sum_{i=1}^{I} \widehat{\beta}_{*i}$ and $t_{(I-1)}$ is the $t-$Student distribution with $I-1$ degrees of freedom.

Proof. Note that $\widehat{\beta}_{*1}, \ldots, \widehat{\beta}_{*I}$ are independent and identically distributed (IID) random variables with $\mathbb{E}(\widehat{\beta}_{*i}) = \beta_*$ and $\mathrm{Var}(\widehat{\beta}_{*i}) = \sigma^2_{\beta_{*i}}$, for $i = 1, \ldots, I$. We must remember the following results of univariate distributions ([33]):

(i) If $\widehat{\beta}_*$ is normally distributed, than $\frac{(I-1)S^2_{\beta_{*i}}}{\sigma^2_{\beta_{*i}}} \sim \chi^2_{(I-1)}$.

(ii) Let X and Y be independent random variables, where $X \sim N(0,1)$ and $Y \sim \chi^2_{(\nu)}$, than $\frac{X}{\sqrt{\frac{Y}{v}}} \sim t_{(\nu)}$.

Now, we can write

$$\frac{\left(\widehat{\boldsymbol{\beta}}_{*i}-\widehat{\widehat{\boldsymbol{\beta}}}_{*}\right)}{S_{\widehat{\beta}_{*i}}} = \frac{\frac{\left(\widehat{\boldsymbol{\beta}}_{*i}-\widehat{\widehat{\boldsymbol{\beta}}}_{*}\right)}{\sigma_{\beta_{*i}}}}{\sqrt{\frac{S^2_{\widehat{\beta}_{*i}}}{\sigma^2_{\beta_{*i}}}}} = \frac{\Delta}{\sqrt{\Gamma}}.$$

The quantities Δ and Γ can be unfolded like:

$$\begin{aligned}\Delta &= \frac{\left(\widehat{\boldsymbol{\beta}}_{*i}-\boldsymbol{\beta}_{*}\right)}{\sigma_{\beta_{*i}}} - \frac{\left(\widehat{\widehat{\boldsymbol{\beta}}}_{*}-\boldsymbol{\beta}_{*}\right)}{\sigma_{\beta_{*i}}} = \frac{\left(\widehat{\boldsymbol{\beta}}_{*i}-\boldsymbol{\beta}_{*}\right)}{\sigma_{\beta_{*i}}} - \frac{1}{\sqrt{I}}\frac{\left(\widehat{\widehat{\boldsymbol{\beta}}}_{*}-\boldsymbol{\beta}_{*}\right)}{\frac{\sigma_{\beta_{*i}}}{\sqrt{I}}}\\ &= C - \frac{1}{\sqrt{I}}D.\end{aligned}$$

The variables C and D are independent and, for large T, $C \dot\sim N(0,1)$ and $D \dot\sim N(0,1)$. Consequently, $\Delta \dot\sim N\left(0, \frac{I+1}{I}\right)$ or $\sqrt{(\frac{I}{I+1})}\Delta \dot\sim N(0,1)$. From (i),

$$\Gamma = \frac{S^2_{\widehat{\beta}_{*i}}}{\sigma^2_{\beta_{*i}}} = \frac{(I-1)S^2_{\widehat{\beta}_{*i}}}{\frac{\sigma^2_{\beta_{*i}}}{(I-1)}} \dot\sim \frac{\chi^2_{(I-1)}}{I-1}.$$

By (ii) it follows that,

$$\frac{\Delta}{\sqrt{\Gamma}} \dot\sim \sqrt{\frac{(I+1)}{I}}t_{(I-1)}, \text{ as } T \text{ increases.}$$

□

Corollary 1. *In case of data coming from a normally distribute variable, the distribution in* (5) *becomes an exact distribution rather than asymptotic one.*

Proof. The same as in Theorem 1. □

Theorem 2. *Let's consider I time series of length T from an ARMA(v,w) process in* (1)*, $\boldsymbol{\beta}$ the vector of parameters defined in* (2) *and $\widehat{\boldsymbol{\beta}}_i$ the vector of OLS estimates for the i^{st} sample, both vectors of length $p = v + w + 1$. Than,*

$$\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right)\boldsymbol{S}^{-1}_{\widehat{\boldsymbol{\beta}}}\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right) \dot\sim \frac{(I-1)(I+1)p}{I(I-p)p}F_{p,I-p}, \text{ as } T \text{ increases,} \quad (6)$$

where $\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}} = \frac{1}{(I-1)}\sum_{i=1}^{I}\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right)\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right)'$ *and* $\widehat{\widehat{\boldsymbol{\beta}}} = \frac{1}{I}\sum_{i=1}^{I}\widehat{\boldsymbol{\beta}}_i$.

Proof. Note that $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_I$ are IID random vectors with $\mathbb{E}(\widehat{\boldsymbol{\beta}}_i) = \boldsymbol{\beta}$ and $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_i) = \Sigma_{\boldsymbol{\beta}}$, for $i = 1, \ldots, I$. We must remember the following results of multivariate distributions ([34]):

(i) If $\widehat{\boldsymbol{\beta}}_i$ is normally distributed, than $(I-1)\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}} \sim \boldsymbol{W}_{p(I-1)}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ ($\boldsymbol{W}$ is the Wishart distribution);

(ii) $\widehat{\boldsymbol{\beta}}_i$ and $\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}}$ are independent;

(iii) It follows from (i) and (ii) that $\left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) \boldsymbol{S}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) \sim \frac{(I-1)p}{I-p} F_{p,(I-1)}$.

Now, let's find the distribution of $(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}})$:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}} &= \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) - \left(\widehat{\widehat{\boldsymbol{\beta}}} - \boldsymbol{\beta}\right) = \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) - \left(\frac{1}{I}\sum_{i=1}^{I} \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) \\
&= \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) - \frac{1}{I}\sum_{i=1}^{I} \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right).
\end{aligned}$$

It follows from (3) that

$$\left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) \dot{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$$

and

$$\frac{1}{I}\sum_{i=1}^{I} \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\right) \dot{\sim} N_p\left(\boldsymbol{0}, \frac{1}{I}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right).$$

.

So,

$$\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right) \dot{\sim} N_p\left(\boldsymbol{0}, \left[\frac{I+1}{I}\right] \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right),$$

or

$$\sqrt{\frac{I}{I+1}} \left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right) \dot{\sim} N_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right).$$

Rewriting (iii) explicitly in terms of probability distributions and considering (3), we have

$$N_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)' \left(\frac{\boldsymbol{W}_{p(I-1)}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}{I-1}\right) N_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \dot{\sim} \frac{(I-1)p}{I-p} F_{p,(I-1)}.$$

Finally, from the equation above we can write (iii) like:

$$\sqrt{\frac{I}{I+1}} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)' \left(\frac{\boldsymbol{W}_{p(I-1)}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}{I-1}\right) \sqrt{\frac{I}{I+1}} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right),$$

or

$$\frac{I}{I+1} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)' \left(\frac{\boldsymbol{W}_{p(I-1)}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}{I-1}\right) N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right).$$

Now we can note that the quantity $\left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right) \boldsymbol{S}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}\right)$ has the following probability distribution:

$$N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)' \left(\frac{\boldsymbol{W}_{p(I-1)}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}{I-1}\right) N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \dot{\sim} \frac{(I+1)}{I} \frac{(I-1)p}{(I-p)} F_{p,(I-1)}, \text{ as } T \text{ increases.}$$

$\square$

Corollary 2. *In case of data coming from a normally distributed variable, the distribution in* (6) *becomes an exact distribution rather than asymptotic one.*

Proof. The same as in Theorem 2. $\square$

2.3. Dynamic ARMA-based control charts

Consider a historical data set of I batches yielding products compliant with specifications. 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). Let's assume that the variable dynamics can be described by the ARMA process.

In order to find a reference distribution of the ARMA (v, w) coefficient estimates in Phase $\mathcal{I}$, we firstly save the OLS vector of estimates $\widehat{\boldsymbol{\beta}}_i$ for each batch. Considering that $E(\widehat{\boldsymbol{\beta}}_i) = \widehat{\boldsymbol{\beta}}$ and $E(\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}}) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, in the next step, we build the unique estimates of the mean and the covariance of $\widehat{\boldsymbol{\beta}}_i$ by combining the individual estimates like:

$$\widehat{\widehat{\boldsymbol{\beta}}} = \frac{1}{I}\sum_{i=1}^{I} \widehat{\boldsymbol{\beta}}_i \quad \text{and} \quad \boldsymbol{S}_{\widehat{\boldsymbol{\beta}}} = \frac{1}{(I-1)} \sum_{i=1}^{I} (\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}})'. \tag{7}$$

These estimates hold relevant information about the variable dynamic (serial correlation) of the process operating in a normal regime. In Phase $\mathcal{II}$ we propose one approach based on the modified *Hotelling* $\mathcal{T}^2$ statistic to monitor the future batch samples. We have shown in Theorem 2 that

$$\mathcal{T}^2_{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}})\boldsymbol{S}^{-1}_{\widehat{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\beta}}_i - \widehat{\widehat{\boldsymbol{\beta}}}) \dot{\sim} \frac{(I-1)(I+1)p}{I(I-p)p} F_{p,I-p}, \tag{8}$$

where $p = v + w + 1$. Scores above the α percentile in $\mathcal{T}^2_{\boldsymbol{\beta}}$ imply that the variable dynamics in a new batch are different to their expected behaviour for the in-control process. Once the $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart pointed out a batch sample out of limit, we can investigate the coefficients most affected by using the control chart based on the modified *t-Student* distribution. We have shown in Theorem 1 that

$$t_{\boldsymbol{\beta}} = \frac{(\widehat{\boldsymbol{\beta}}_{*i} - \widehat{\widehat{\boldsymbol{\beta}}}_*)}{S_{\widehat{\beta}_{*i}}} \dot{\sim} \sqrt{\frac{(I+1)}{I}} t_{(I-1)}, \tag{9}$$

where $\widehat{\boldsymbol{\beta}}_{*i}$ and $\widehat{\widehat{\boldsymbol{\beta}}}_*$ are elements of the vectors $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\widehat{\boldsymbol{\beta}}}$, respectively. $S^2_{\widehat{\beta}_{*i}}$ is an element of the main diagonal of $\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}}$. Scores above the α percentile in $t_{\boldsymbol{\beta}}$ can signalize a disturbance in the dynamic caused by the changing in a specific coefficient.

3. Simulation study

In this Section, we generate batch processes in which the dynamic is described by an ARMA(v,w) model. In order to illustrate our method, we present a Monte Carlo simulation using varieties of this model, including combinations of $v, w = 0, 1, 2$. Both models with intercept term. The model with the highest number of parameters in this study is an ARMA(2,2), explicitly written as

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} \tag{10}$$

with the vector of parameters $\boldsymbol{\beta} = [\phi_0, \phi_1, \phi_2, \theta_1, \theta_2]$. Table 1 show the set of ARMA parameters for in-control process and the simulation settings. In phase $\mathcal{II}$, we considering scenarios with a wide range of disturbances in the intercept term ϕ_0 and in the AR/MA part of the model, represented by ϕ_1 and θ_1, respectively.

We generate scenarios including different numbers of batches with different time-length (from 100 to 1000). Each scenario was replicated 1000 times. In phase $\mathcal{I}$ we do variate the number of batches from 30 to 100. The $\mathcal{T}^2_{\boldsymbol{\beta}}$ (8) and $t_{\boldsymbol{\beta}}$ (9) charts were setting to the false alarm probability of $\alpha = 0.01$. In phase $\mathcal{II}$ 500 batches were generated in each scenario. The rate of batches beyond the control (r) and the ARL index (*Average Run Length*) 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 $\alpha = 1\%$, $ARL_0 = 100$), i.e., points above control

Table 1: Simulation settings

ARMA coefficients	ARMA settings	Disturbed parameter	Disturbance levels	# Batches phase $\mathcal{I}$	# Batches phase $\mathcal{II}$	Batch length T	Run
ϕ_0	1	X	$0, 0.5, 0.8, 1, 1.2, 1.5, 2$	30, 50, 100	500	100, 200, 500, 1000	1000
ϕ_1	0.2	X	$-0.2, 0, 0.1, 0.2, 0.3$				
ϕ_2	0.5						
θ_1	0.5	X	$0, 0.3, 0.4, 0.5, 0.6, 0.8$				
θ_2	-0.3						

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.

As a benchmark approach, we consider the usual way to build time series-based control charts for monitoring the residuals from the fitted model in phase $\mathcal{I}$ [32] by adapting this methodology for the case of batch processes. We use as the variable the residual mean $\overline{e}$ from the $T-p$ residuals for each batch, where $p = v + w + 1$ for the ARMA(v, w) model with intercept. We know that if a new batch comes from the in-control process, $\overline{e} \dot{\sim} N(0, 1/\sqrt{T-p})$. We set the limits of the $t_{\boldsymbol{e}}$ residual control chart using the probability of false alarm of $\alpha = 1\%$. The EWMA approach [32] is used in order to improve the power of the $t_{\boldsymbol{e}}$ chart to detect disturbances representing small changes in the residual mean. Simulations and calculations were conducted using R [35].

Tables 2 to 4 summarize the results of $\mathcal{T}^2_{\boldsymbol{\beta}}$ and $t_{\boldsymbol{e}}$ charts for an ARMA(1,1) model with the in-control parameters set in Table 1. The tables show the mean and standard deviation of ARL values for each disturbance. The scenarios in Tables 2 and 3 are very similar in terms of results and so they will not be commented apart. These Tables include disturbances in AR coefficient ϕ_1 and MA coefficient θ_1, each one at the time. The observed ARL_0 (highlighted in the gray line) is close to the chosen nominal value of 100, which is consistent with the fact that no disturbance was introduced in the process. The ARL_1 values show that the $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart outperforms the $t_{\boldsymbol{e}}$ in detecting disturbances of different intensities. Additionally, we notice that the degree of detection in $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart increases faster as the perturbations get more intense. Even for the higher values of disturbances, the performance of our approach remains better than the residual-based charts. We emphasize here the power of the proposed approach (based on estimates of correlations in ϕ) to capture information about process dynamics.

Table 4 shows the in-control process based on the ARMA(1,1) model and disturbances included in the intercept parameter ϕ_0, in which represent levels of change in the process mean, i.e., those time series are

Table 2: ARMA(1,1): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the AR coefficient ϕ_1

ϕ_1	T	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$
-0.2	100	1.55	0.41	207.10	198.22	1.37	0.23	261.99	206.90	1.32	0.12	301.52	195.96
	200	1.03	0.04	108.04	136.28	1.02	0.02	245.43	190.74	1.01	0.01	340.91	159.01
	500	1.00	0.00	206.38	191.31	1.00	0.00	213.12	176.74	1.00	0.00	272.73	154.07
	1000	1.00	0.00	235.10	197.34	1.00	0.00	244.09	199.38	1.00	0.00	315.15	186.85
0.0	100	10.47	7.85	198.35	184.84	7.69	3.47	223.94	183.48	6.67	2.35	242.15	184.06
	200	3.88	2.66	158.03	164.89	2.89	0.97	224.23	184.59	2.64	0.59	317.85	174.84
	500	1.24	0.20	142.02	161.94	1.17	0.11	208.81	167.39	1.11	0.05	320.03	192.10
	1000	1.00	0.01	160.56	168.54	1.00	0.01	210.58	188.11	1.00	0.00	292.56	183.92
0.1	100	50.61	53.17	163.09	170.65	30.24	19.69	169.52	151.33	29.49	20.40	223.92	177.65
	200	37.62	39.76	136.60	136.04	19.03	11.79	176.14	162.34	16.92	8.64	219.03	164.46
	500	8.31	6.87	158.23	169.50	7.07	7.29	164.81	150.12	4.75	1.42	268.94	183.23
	1000	2.88	1.67	193.80	193.80	2.19	0.86	185.32	171.76	1.92	0.32	236.66	170.04
0.2	100	143.14	149.41	116.17	144.96	114.40	107.31	132.28	153.58	106.96	97.14	182.32	168.63
	200	154.88	146.54	125.29	160.91	129.26	112.50	111.44	129.12	121.95	109.16	180.62	152.30
	500	198.93	167.21	116.80	150.20	158.91	145.74	171.53	174.05	132.75	116.87	179.84	150.86
	1000	145.57	139.15	121.50	157.18	195.10	160.02	152.59	155.07	149.26	127.64	190.92	166.38
0.3	100	89.42	123.93	74.85	113.05	64.15	86.76	104.44	125.94	36.17	21.05	88.44	111.81
	200	65.90	110.76	66.32	100.31	32.98	52.26	101.62	120.87	20.54	10.19	104.85	129.28
	500	10.05	7.85	94.77	135.57	6.66	3.24	95.23	126.16	5.51	1.94	89.50	113.80
	1000	2.91	1.58	99.17	123.37	2.29	0.78	118.02	140.85	1.92	0.37	115.91	133.89
0.6	100	1.46	0.44	7.20	5.67	1.30	0.21	7.26	4.08	1.23	0.12	6.77	2.66
	200	1.02	0.03	6.68	4.59	1.01	0.01	6.98	3.81	1.00	0.00	6.72	2.42
	500	1.00	0.00	5.76	2.33	1.00	0.00	6.63	2.90	1.00	0.00	7.03	2.68
	1000	1.00	0.00	7.52	6.95	1.00	0.00	8.04	10.00	1.00	0.00	7.46	3.39

generated from different mean drifts. The $t_{\boldsymbol{e}}$ is well suitable to capture this kind of change, as expected. Even being considered as the *underdog* in this scenario, we noticed the $\mathcal{T}^2_{\boldsymbol{\beta}}$ performance close to the residual one for the highest number of reference batches with the highest time-instants in any disturbances. In other words, even for the changes in the mean, instead of in the data dynamic, our approach seems to work well.

Tables S1 to S4 in the Supplementary Material show the results of Monte Carlo Simulations for ARMA(2,2), AR(1) and MA(1), respectively. They are very similar compared to the study presented in Tables 2 to 4.

Table 3: ARMA(1,1): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the MA coefficient θ_1

θ_1	T	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\boldsymbol{e}}(\hat{\mu})$	$t_{\boldsymbol{e}}(\hat{\sigma})$
0.0	100	1.07	0.07	229.47	211.46	1.05	0.03	231.92	197.23	1.04	0.02	287.27	187.26
	200	1.00	0.00	257.31	193.96	1.00	0.00	171.38	186.29	1.00	0.00	359.66	190.02
	500	1.00	0.00	218.69	189.65	1.00	0.00	276.63	208.32	1.00	0.00	272.92	149.85
	1000	1.00	0.00	257.65	223.60	1.00	0.00	314.74	180.34	1.00	0.00	427.78	148.14
0.3	100	6.23	3.60	161.44	165.48	5.21	1.95	201.47	177.20	4.65	1.25	241.61	184.57
	200	2.58	1.13	153.24	181.85	2.25	0.67	216.95	194.58	2.00	0.32	316.91	188.82
	500	1.12	0.10	161.22	164.04	1.07	0.06	203.95	177.16	1.05	0.03	288.90	186.49
	1000	1.00	0.00	176.50	174.51	1.00	0.00	217.00	189.18	1.00	0.00	289.73	188.91
0.4	100	26.83	20.14	162.86	187.89	24.73	15.73	195.65	182.32	20.11	8.02	238.83	172.32
	200	17.30	15.87	148.62	176.75	13.48	8.29	192.74	178.25	11.14	5.23	254.78	191.96
	500	6.04	4.39	137.97	160.31	4.05	2.01	207.24	182.84	3.44	0.88	188.90	138.53
	1000	2.03	0.87	145.61	165.84	1.67	0.33	181.28	171.34	1.53	0.22	219.33	167.85
0.5	100	143.14	149.41	116.17	144.96	114.40	107.31	132.28	153.58	106.96	97.14	182.32	168.63
	200	154.88	146.54	125.29	160.91	129.26	112.50	111.44	129.12	121.95	109.16	180.62	152.30
	500	198.93	167.21	116.80	150.20	158.91	145.74	171.53	174.05	132.75	116.87	179.84	150.86
	1000	145.57	139.15	121.50	157.18	195.10	160.02	152.59	155.07	149.26	127.64	190.92	166.38
0.6	100	76.74	102.94	96.58	126.08	59.45	61.19	89.62	116.01	46.27	54.90	127.97	139.03
	200	46.68	80.02	96.13	129.03	24.11	21.61	97.20	117.70	20.85	12.69	170.86	165.75
	500	11.66	49.70	99.45	127.41	4.95	2.69	117.78	129.75	4.05	1.21	158.33	154.58
	1000	2.09	1.18	105.79	153.59	1.69	0.48	102.90	127.67	1.56	0.27	163.78	159.80
0.8	100	2.80	1.70	70.04	102.65	2.16	0.72	70.07	88.11	2.00	0.47	75.62	98.99
	200	1.12	0.19	64.76	100.40	1.08	0.08	88.80	116.40	1.04	0.03	92.68	108.22
	500	1.00	0.00	67.82	105.03	1.00	0.00	80.83	132.12	1.00	0.00	80.54	107.80
	1000	1.00	0.00	80.70	127.66	1.00	0.00	71.50	99.24	1.00	0.00	97.54	111.18

Table 4: ARMA(1,1): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in in the intercept ϕ_0

ϕ_0	T	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$
0.0	100	1.07	0.08	1.00	0.00	1.04	0.03	1.00	0.00	1.02	0.01	1.00	0.00
	200	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	500	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	1000	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
0.5	100	4.93	3.12	1.00	0.00	3.89	1.63	1.00	0.00	3.12	0.76	1.00	0.00
	200	1.91	0.74	1.00	0.00	1.52	0.32	1.00	0.00	1.44	0.23	1.00	0.00
	500	1.02	0.03	1.00	0.00	1.01	0.01	1.00	0.00	1.00	0.00	1.00	0.00
	1000	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
0.8	100	63.39	77.90	3.85	8.47	38.90	32.02	2.41	1.86	30.80	16.68	2.30	1.62
	200	40.05	74.13	1.26	0.94	20.85	16.44	1.14	0.17	16.60	8.25	1.13	0.13
	500	8.68	7.46	1.01	0.04	5.22	2.15	1.01	0.01	4.63	1.51	1.00	0.00
	1000	2.61	1.48	1.00	0.00	1.93	0.55	1.00	0.00	1.75	0.29	1.00	0.00
1.0	100	163.85	158.72	143.12	165.64	107.67	90.98	156.41	171.88	94.23	78.88	176.62	159.59
	200	141.20	116.93	91.45	135.41	141.52	118.74	158.14	157.23	116.36	89.73	208.97	178.09
	500	174.39	151.85	125.42	158.62	154.68	135.68	161.39	178.08	120.63	90.08	207.97	181.85
	1000	175.58	158.47	118.26	158.62	164.63	129.10	167.33	167.43	138.47	105.63	207.66	180.26
1.2	100	66.71	91.16	2.81	3.92	41.20	34.88	2.00	2.42	31.48	14.39	1.52	0.45
	200	44.10	76.60	1.13	0.27	22.69	19.12	1.08	0.09	15.15	6.78	1.08	0.11
	500	7.20	7.47	1.00	0.00	5.25	2.64	1.00	0.00	4.53	1.66	1.00	0.00
	1000	2.55	1.51	1.00	0.00	1.94	0.52	1.00	0.00	1.76	0.33	1.00	0.00
1.5	100	5.81	4.59	1.00	0.00	4.11	1.67	1.00	0.00	3.33	0.98	1.00	0.00
	200	1.83	0.86	1.00	0.00	1.62	0.44	1.00	0.00	1.38	0.19	1.00	0.00
	500	1.02	0.03	1.00	0.00	1.01	0.01	1.00	0.00	1.00	0.00	1.00	0.00
	1000	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
2.0	100	1.07	0.06	1.00	0.00	1.03	0.02	1.00	0.00	1.02	0.01	1.00	0.00
	200	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	500	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	1000	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

3.1. Monitoring and diagnosis case

This Section is aimed to show the full potential of our approach. For a simulated ongoing batch process we display the group of T^2_β chart and t_β charts for the individual coefficients in order to show its performance to detect and diagnose the imposed disturbances.

Let's consider the industrial process generating batches of 200 time-length following the ARMA(1,1) model, with ϕ_0, ϕ_1 and θ_1 set as in Table 1. We built the charts using $\alpha = 0.01$ and considered 30 in-control batches as the reference. We simulate 20 new batches with two levels of disturbances in the ϕ_1 parameter: (i) a moderate level (from ϕ_1=0.2 to ϕ_1=0); and (ii) a intense level (from ϕ_1=0.2 to ϕ_1=0.6).

Figures 1 and 2 show the group of control charts for the moderate and intense level of disturbance, respectively. In Figure 1 we noticed that the multivariate T^2_β starts to signalize a number of points out of the limits just after the moderate disturbance has imposed (after the 30^{th} batch). It reinforces a good performance in the simulated study shown in Tables 2 and 3, as expected. Additionally, from the t_β charts there is a good tip about the source of the disturbance, since the t_β chart for ϕ_1 points out some points beyond the limits. The other two t_β charts remain with nearly all points randomly running within the charts boundaries, as it should be, since there are no disturbances imposed in ϕ_0 and θ_1.

In Figure 2 it becomes even more pronounced since the level of disturbance is higher. We can see nearly all points out of limits in the T^2_β and t_β chart for ϕ_1. The t_β for ϕ_0 show a few points outside the limits after the 30^{th} batch. Those few false alarms are likely to be due to coefficient covariance. In general, these charts seem to work really well to signalize and diagnose the source of disturbance.

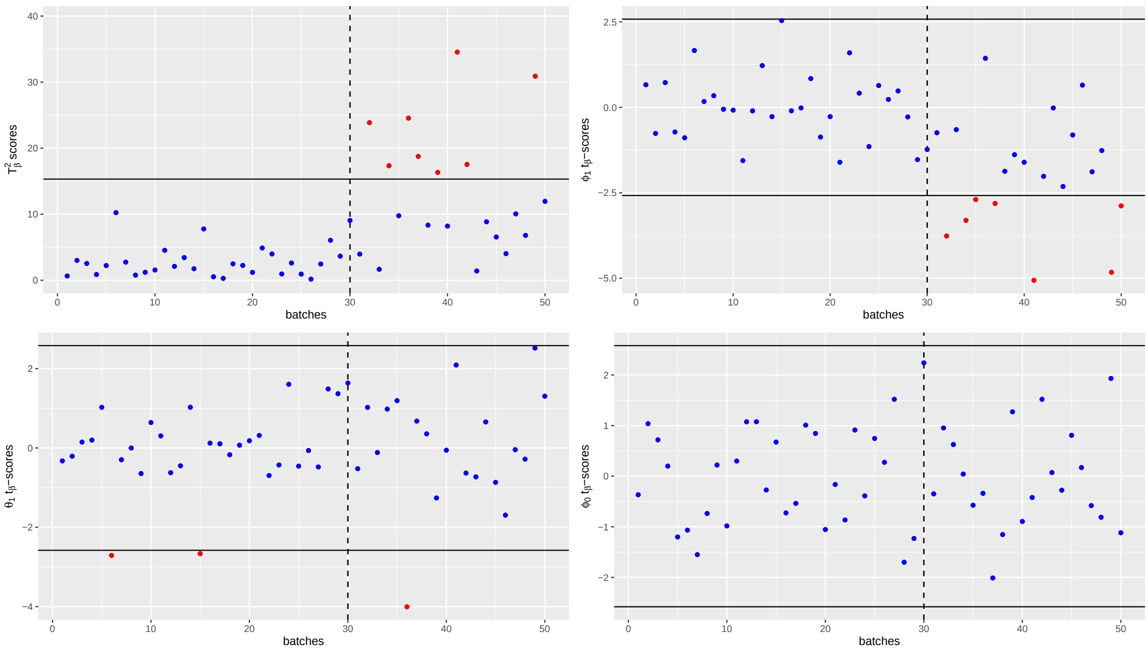

Figure 1: T_β^2 and t_β for 30 in-control batches and 20 new batches with a change in ϕ_1 from 0.2 to 0.

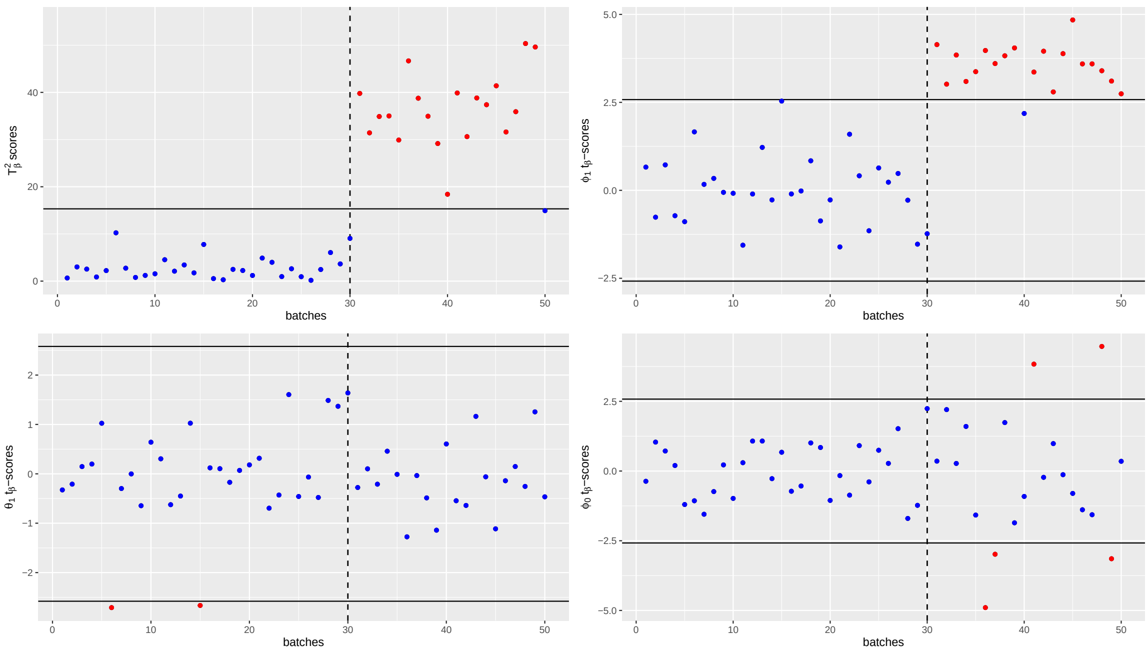

Figure 2: T_β^2 and t_β for 30 in-control batches and 20 new batches with a change in ϕ_1 from 0.2 to 0.6.

4. Application

In order to illustrate the applicability of our methodology we consider a real dataset of time series representing measurements of engine noise from [36]. We can understand each time series as one batch sampled from an industrial process. Let's consider only the training dataset which has 3271 batches, each one with 500 time instants, sampled from the process operating under two different conditions, labeled as +1 and -1 with 1755 and 1846 number of batches, respectively. In this application we assume that the group labeled as +1 is the reference group, i.e., sample batches coming from the process operating in a standard condition.

The chart on the left on Figure (3) shows the average behaviour of the autocorrelation function (ACF) for all batches according to their group. The colours green and red represent the reference group (labeled as +1) and the monitoring group (labeled as -1), respectively. We note that the main feature is the cyclical behavior of data in both groups. Following the aim of our methodology, in Phase $\mathcal{I}$ data from the reference group are modeled by using the class of ARMA models. We know that this class of models is suitable to capture the time series dynamic rather than other features like cycles, trends, etc. For that reason we chose an AR model of high order to capture the cycles as the dynamics. Here an AR(12) was adopted in order to model this feature since the order 12 was necessary and sufficient for getting uncorrelated and normally distributed residuals. The Ljung–Box test pointed out that 100% of the residuals are uncorrelated, non significant to presence of correlation and 95% of the residuals were normally distributed according to the Shapiro-Wilk test for normality, using a significance level of 5%.

Although we have 12 AR coefficients in the model for the reference group, we noticed in Figure (3) on the right chart that the large portion of them is possibly not significantly different from zero. Thus, proceeding an individual t-test for each coefficient and verifying their statistical significance at a level of 5%, only the first three coefficients of the fitted AR model were significant in 95% of batches. These results are summarized in Table 5. For this reason, the $\mathcal{T}^2_{\boldsymbol{\beta}}$ control chart was built with the first three coefficients. Its seems to be a good choice insofar as Figure (3) shows a clear visual difference between the means of the adjusted coefficients from group +1 and -1 just in those 3 first AR coefficients (represented by green and red lines).

Table 5: Rate of significant coefficients in the individual t-test for the reference group adjusted AR(12) model.

# Batches	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}	ϕ_{11}	ϕ_{12}
1755	1.00	1.00	0.98	0.35	0.48	0.20	0.30	0.20	0.18	0.25	0.29	0.51

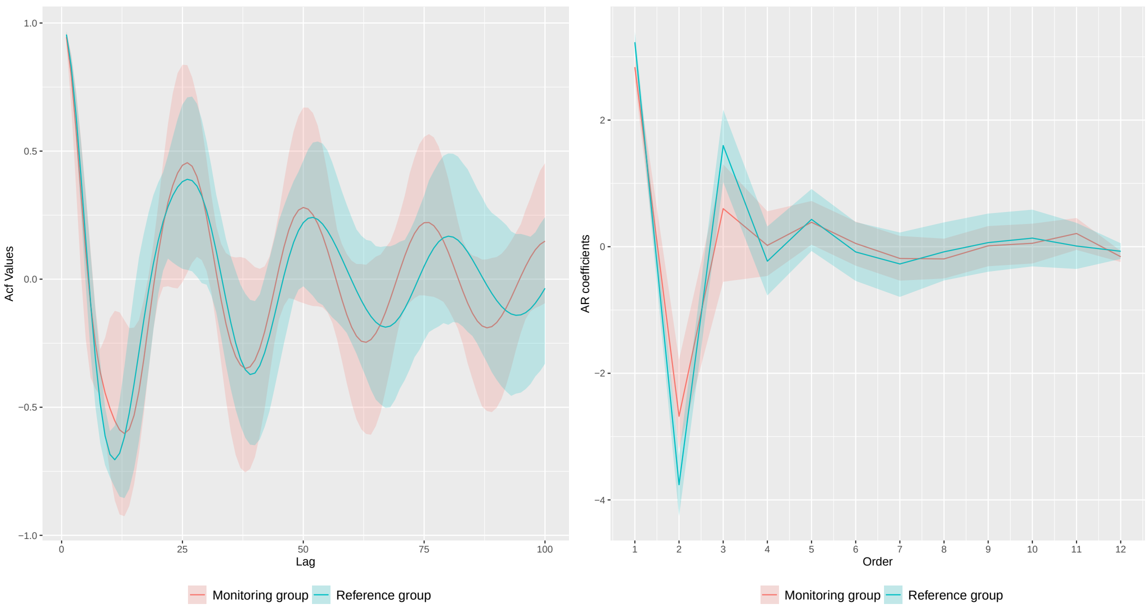

Figure 3: Autocorrelations for the FordA-train dataset and AR coeficients from the ajusted AR(12) model for all batches.

In order to show the performance of our approach we take randomly I in-control batches (i.e., from the ones labeled as +1) as the reference batches to fit the AR(12) model and build the $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart. We do variate the study by using I values of 10, 30, 100, 200, 300, 500 with 200 replications each. The control limits are set with false alarm probability (α) of 0.10, 0.05 and 0.01. The remaining 1755 - I batches are used to evaluate the empirical false alarm probability. The 1846 batches labeled as -1 are used to evaluate the power of $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart.

Table 6 summarize the results of $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart. The r_0 and r_1 are the rate of false alarm and disturbed batches detected, respectively. We noticed that the observed r_0 (highlighted in the gray line) is closer to the chosen nominal value of α as the number of samples increases, which is consistent with the theoretical distribution derived in Theorem 2. The r_1 values show the performance of $\mathcal{T}^2_{\boldsymbol{\beta}}$ chart to signalize the out-of-control (labeled as -1) batches. As we expected the degree of detection increases as the number of reference batches I in phase $\mathcal{I}$ increases. Even for the very small number of batches compared to the overall number of batches available labeled as +1, the $\mathcal{T}^2_{\boldsymbol{\beta}}$ shows a good rate of detection for each false alarm probability α.

Table 6: AR(12): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of r_0 (in gray) and r_1 values

		False alarm probability (α)					
		0.10		0.05		0.01	
	I	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$
Phase $\mathcal{I}$	10	0.13	0.10	0.08	0.08	0.03	0.04
	100	0.10	0.02	0.07	0.02	0.03	0.01
	200	0.10	0.02	0.07	0.01	0.03	0.01
	300	0.09	0.01	0.06	0.01	0.03	0.01
	500	0.08	0.01	0.05	0.01	0.02	0.01
Phase $\mathcal{II}$	10	0.76	0.17	0.61	0.22	0.31	0.22
	30	0.87	0.07	0.79	0.09	0.57	0.15
	100	0.89	0.03	0.83	0.04	0.67	0.07
	200	0.89	0.02	0.83	0.03	0.68	0.04
	300	0.89	0.02	0.84	0.02	0.69	0.03
	500	0.89	0.01	0.84	0.02	0.70	0.03

5. Conclusion

This paper introduced a new approach to deal with batch processes through a set of ARMA-based control charts. Through in-control batch samples available we fitted the ARMA model and built a group of charts based on the coefficient estimates from historical in-control batches. The modified *Hotelling* and *t-Stutent* distributions can easily accommodate those estimates and a decision rule was made for monitoring future samples. Additionally, the modified t-Student charts help to look for the source of disturbances.

The simulated batch process generating samples of time series from an ARMA model was presented. The T^2_{beta} chart outperforms the traditional competitor based on the residuals for detecting changes of any level in the process dynamic. Furthermore, we have shown how powerful are the individual t_β charts to identify the source of disturbances imposed in the process.

The applicability of our approach was illustrated through a real data set in which the good performance is clearly noticed, even for a very small sample reference batches compared to the overall number of in-control batches available.

Finally, it's important to noticed that we can built the group of T^2_{beta} and t_β charts from any ARMA (v,w) sub model, including only AR(v) or MA(w) component with the order less or equal to v or w, with and without intercept. It opens the applicability of this approach to a wide range of batch processes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material: Supplementary tables.

Code: R-functions containing all methods developed in this article (will be available in the dvqcc package at CRAN).

Data: Dataset used in the application and corresponding script (zip).

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] T.-H. Kuang, Z. Yan, Y. Yao, Multivariate fault isolation via variable selection in discriminant analysis, Journal of Process Control 35 (2015) 30–40.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. Snoussi, Spc for short-run multivariate autocorrelated processes, Journal of Applied Statistics 38 (10) (2011) 2303–2312.

[13] X. Pan, J. E. Jarrett, Why and how to use vector autoregressive models for quality control: the guideline and procedures, Quality & Quantity 46 (3) (2012) 935–948.

[14] E. Vanhatalo, M. Kulahci, The effect of autocorrelation on the hotelling t2 control chart, Quality and Reliability Engineering International 31 (8) (2015) 1779–1796.

[15] R. C. Leoni, A. F. B. Costa, B. C. Franco, M. A. G. Machado, The skipping strategy to reduce the effect of the autocorrelation on the t2 chart's performance, The International Journal of Advanced Manufacturing Technology 80 (9-12) (2015) 1547–1559.

[16] F. D. Simões, R. C. Leoni, M. A. G. Machado, A. F. B. Costa, Synthetic charts to control bivariate processes with autocorrelated data, Computers & Industrial Engineering 97 (2016) 15–25.

[17] L. C. Alwan, H. V. Roberts, Time-series modeling for statistical process control, Journal of business & economic statistics 6 (1) (1988) 87–95.

[18] W. Jiang, K.-L. Tsui, W. H. Woodall, A new spc monitoring method: The arma chart, Technometrics 42 (4) (2000) 399–410.

[19] W. Jiang, K. Tsui, Some properties of the arma control chart, NONLINEAR analysis-theory METHODS & APPLICATIONS 47 (3) (2001) 2073–2088.

[20] W. Jiang, Average run length computation of arma charts for stationary processes, Communications in Statistics-Simulation and Computation 30 (3) (2007) 699–716.

[21] H. De la Torre Gutiérrez, D. T. Pham, Identification of patterns in control charts for processes with statistically correlated noise, International Journal of Production Research 56 (4) (2018) 1504–1520.

[22] H. De la Torre-Gutiérrez, D. Pham, A control chart pattern recognition system for feedback-control processes, Expert Systems with Applications 138 (2019) 112826.

[23] A. Costa, S. Fichera, Economic-statistical design of adaptive arma control chart for autocorrelated data, Journal of Statistical Computation and Simulation 91 (3) (2021) 623–647.

[24] R. Sheikhrabori, M. Aminnayeri, Maximum likelihood estimation of the change point in stationary state of auto regressive moving average (arma) models, using svd-based smoothing, Communications in Statistics-Theory and Methods (2021) 1–19.

[25] H. Nguyen, A. A. Nadi, K. Tran, P. Castagliola, G. Celano, K. Tran, The effect of autocorrelation on the shewhart-rz control chart, arXiv preprint arXiv:2108.05239 (2021).

[26] S. Jafarian-Namin, M. S. Fallahnezhad, R. Tavakkoli-Moghaddam, A. Salmasnia, S. M. T. Fatemi Ghomi, An integrated quality, maintenance and production model based on the delayed monitoring under the arma control chart, Journal of Statistical Computation and Simulation (2021) 1–25.

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

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

[29] 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.

[30] G. E. Box, G. M. Jenkins, G. Reinsel, et al., Forecasting and control, Time Series Analysis 3 (1970) 75.

[31] P. J. Brockwell, R. A. Davis, Stationary time series, in: Time Series: Theory and Methods, Springer, 1991, pp. 1–41.

[32] D. C. Montgomery, Introduction to statistical quality control, John Wiley & Sons, 2007.

[33] A. Mood, F. Graybill, D. Boes, (1974), introduction to the theory of statistics (1974).

[34] R. A. Johnson, D. W. Wichern, et al., Applied multivariate statistical analysis, Vol. 5, Prentice hall Upper Saddle River, NJ, 2002.

[35] 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/`

[36] Time-Series-Classification-Website, Time series classification website, `http://www.timeseriesclassification.com/description.php?Dataset=FordA`, (Accessed on 07/08/2021).

Supplementary Material
Fault detection and diagnosis of batch process using dynamic ARMA-based control charts

Oliveira, B., Valk, M. and Marcondes Filho, D.

May 2021

S1 Simulation studies, ARL results

Table S1: ARMA(2,2): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the AR coefficient ϕ_1

ϕ_1	n	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$
-0.2	100	1.75	0.68	119.08	132.07	1.42	0.37	170.16	166.18	1.28	0.12	240.64	195.06
	200	1.03	0.04	182.76	194.86	1.01	0.02	187.33	182.39	1.01	0.01	247.15	169.02
	500	1.00	0.00	161.90	171.18	1.00	0.00	168.69	157.86	1.00	0.00	265.97	175.27
	1000	1.00	0.00	167.28	174.64	1.00	0.00	262.81	205.03	1.00	0.00	274.49	185.28
0.0	100	19.30	21.56	120.67	156.44	12.32	7.28	121.34	143.25	8.25	2.91	169.26	176.95
	200	5.11	5.14	130.13	163.29	3.31	2.06	191.75	171.62	2.65	0.67	190.58	156.92
	500	1.28	0.37	151.35	176.47	1.13	0.14	217.78	194.64	1.08	0.05	262.67	175.65
	1000	1.00	0.01	169.44	177.39	1.00	0.00	201.36	192.28	1.00	0.00	237.27	176.67
0.1	100	69.64	87.56	109.75	155.72	39.59	24.01	120.97	150.82	34.66	19.33	162.23	164.99
	200	42.73	62.94	136.78	169.31	26.53	16.55	140.70	157.57	20.01	11.24	161.88	163.29
	500	10.99	9.34	134.83	152.74	7.15	3.87	183.58	177.94	5.09	2.21	241.68	186.36
	1000	3.29	2.22	156.76	178.81	2.18	0.76	187.01	173.28	1.90	0.49	213.31	165.24
0.2	100	108.41	106.26	116.21	160.89	69.67	72.84	98.18	136.31	57.74	33.47	131.83	144.95
	200	147.25	131.61	107.03	148.55	101.11	93.56	151.25	166.54	81.21	47.38	154.94	152.91
	500	195.45	173.34	136.17	167.56	141.81	120.70	151.72	150.83	123.25	105.34	193.01	173.96
	1000	159.28	136.46	110.93	141.98	145.53	136.01	164.96	177.67	159.60	135.72	190.66	166.11
0.3	100	47.74	43.79	82.80	128.95	38.62	33.69	83.95	126.42	27.64	13.86	69.28	82.46
	200	35.86	54.90	87.13	115.27	19.97	15.51	90.45	118.87	16.50	7.04	105.78	116.27
	500	12.00	12.75	96.27	140.94	6.56	3.67	120.16	149.62	4.81	1.42	143.98	150.88
	1000	2.95	1.42	94.71	128.58	2.13	0.69	111.22	132.25	1.78	0.29	148.01	154.21
0.6	100	1.45	0.41	17.91	22.82	1.30	0.21	19.48	29.99	1.20	0.10	23.46	49.44
	200	1.02	0.03	21.65	22.19	1.01	0.01	24.03	24.24	1.00	0.00	18.88	14.19
	500	1.00	0.00	19.86	17.90	1.00	0.00	33.38	70.19	1.00	0.00	25.62	29.91
	1000	1.00	0.00	35.49	63.47	1.00	0.00	26.32	50.90	1.00	0.00	25.96	17.42

Table S2: ARMA(2,2): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the MA coefficient θ_1

θ_1	n	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$
0.0	100	1.11	0.10	144.51	175.89	1.08	0.05	256.98	190.28	1.05	0.02	268.03	187.72
	200	1.00	0.00	116.11	131.36	1.00	0.00	353.39	194.25	1.00	0.00	393.06	178.35
	500	1.00	0.00	175.38	202.64	1.00	0.00	319.53	184.33	1.00	0.00	125.00	34.02
	1000	1.00	0.00	185.68	187.93	1.00	0.00	185.28	178.41	1.00	0.00	444.44	136.08
0.3	100	8.00	4.99	136.29	160.20	6.26	3.06	182.97	186.39	4.76	1.48	197.76	175.05
	200	2.70	1.20	123.73	157.39	1.97	0.60	209.82	196.41	1.75	0.28	243.51	190.46
	500	1.05	0.05	132.25	159.16	1.03	0.03	262.03	191.05	1.02	0.01	259.94	187.60
	1000	1.00	0.00	169.35	180.83	1.00	0.00	220.45	180.69	1.00	0.00	260.86	181.08
0.4	100	35.63	38.49	141.76	180.17	22.84	15.75	118.50	144.94	20.44	10.18	136.59	153.95
	200	23.81	50.30	131.90	161.53	12.34	7.20	160.64	170.64	9.32	3.32	166.07	160.66
	500	4.08	3.08	151.65	154.83	3.00	1.27	172.15	174.77	2.48	0.59	217.40	167.94
	1000	1.52	0.40	148.86	173.68	1.30	0.18	182.55	175.44	1.23	0.10	248.07	193.93
0.5	100	108.41	106.26	116.21	160.89	69.67	72.84	98.18	136.31	57.74	33.47	131.83	144.95
	200	147.25	131.61	107.03	148.55	101.11	93.56	151.25	166.54	81.21	47.38	154.94	152.91
	500	195.45	173.34	136.17	167.56	141.81	120.70	151.72	150.83	123.25	105.34	193.01	173.96
	1000	159.28	136.46	110.93	141.98	145.53	136.01	164.96	177.67	159.60	135.72	190.66	166.11
0.6	100	72.77	82.80	51.91	74.90	53.05	58.34	60.14	76.11	43.93	52.68	102.80	138.96
	200	43.21	65.75	84.29	119.22	22.45	21.86	92.66	115.89	15.26	7.92	122.13	139.17
	500	5.97	6.16	120.76	145.44	3.42	2.00	98.38	123.92	2.68	1.06	146.43	142.70
	1000	1.46	0.63	83.24	110.53	1.25	0.25	87.16	100.96	1.11	0.09	117.72	126.48
0.8	100	33.22	32.29	46.57	71.46	20.68	13.60	38.71	64.99	16.18	6.36	34.45	34.13
	200	12.73	13.40	60.30	95.53	7.73	4.59	68.52	113.41	6.31	2.39	49.18	75.96
	500	2.33	1.61	61.21	103.43	1.78	0.58	56.66	95.45	1.42	0.24	57.64	78.97
	1000	1.07	0.12	52.02	93.04	1.02	0.03	61.19	103.85	1.01	0.01	54.06	62.01

Table S3: ARMA(1,0): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the AR coefficient ϕ_1

ϕ_1	n	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_e(\hat{\mu})$	$t_e(\hat{\sigma})$
-0.2	100	1.46	0.40	166.89	177.24	1.31	0.15	282.67	200.61	1.22	0.08	310.88	195.12
	200	1.02	0.05	196.16	196.63	1.01	0.01	223.91	174.06	1.01	0.01	342.26	199.77
	500	1.00	0.00	165.65	178.92	1.00	0.00	260.15	191.40	1.00	0.00	335.42	180.48
	1000	1.00	0.00	152.55	160.56	1.00	0.00	191.39	185.40	1.00	0.00	286.26	212.32
0.0	100	12.20	11.08	180.42	171.12	9.59	6.93	236.21	202.75	6.87	2.55	299.07	196.04
	200	3.65	2.45	177.43	191.51	2.57	0.85	230.59	198.37	2.35	0.58	259.12	184.67
	500	1.16	0.14	191.53	191.77	1.12	0.09	214.12	176.88	1.09	0.05	356.81	163.24
	1000	1.00	0.00	206.25	187.12	1.00	0.00	228.02	164.90	1.00	0.00	238.42	179.93
0.1	100	90.92	121.05	149.02	169.22	52.69	52.07	179.04	185.44	47.48	69.66	211.54	171.56
	200	27.59	25.98	122.76	134.66	19.66	17.78	230.68	192.85	17.52	10.29	256.31	182.58
	500	6.93	6.70	169.45	177.25	4.85	2.14	190.27	180.37	4.11	1.26	260.13	185.81
	1000	2.02	0.87	132.94	146.15	1.93	0.49	205.34	185.44	1.74	0.25	254.19	175.22
0.2	100	208.81	159.50	96.07	122.54	167.91	142.16	114.85	139.49	165.49	128.39	159.95	156.46
	200	189.09	163.23	109.35	130.74	181.86	141.37	127.76	138.40	136.00	113.99	147.42	152.84
	500	187.67	157.67	104.11	141.36	156.56	132.89	140.00	142.87	150.90	115.36	182.02	166.72
	1000	156.78	137.82	126.97	154.74	187.34	154.93	121.61	138.28	147.10	104.13	194.75	161.37
0.3	100	59.57	81.81	82.55	125.40	37.66	32.65	80.51	110.81	27.86	13.15	88.38	110.10
	200	30.03	57.46	72.16	113.80	16.03	10.99	90.91	112.08	14.82	6.84	128.45	133.24
	500	5.73	3.80	53.64	74.44	5.00	2.13	93.37	119.82	3.90	1.04	114.12	134.84
	1000	2.12	0.93	116.96	153.30	1.84	0.49	80.10	107.45	1.63	0.21	98.27	112.39
0.6	100	1.22	0.15	7.23	5.22	1.13	0.07	6.51	3.22	1.11	0.05	6.54	2.59
	200	1.01	0.01	6.80	5.09	1.00	0.00	6.89	4.04	1.00	0.00	7.14	2.71
	500	1.00	0.00	7.21	4.72	1.00	0.00	7.62	3.61	1.00	0.00	7.24	3.45
	1000	1.00	0.00	8.29	6.11	1.00	0.00	7.07	3.29	1.00	0.00	6.89	2.36

Table S4: ARMA(0,1): Mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of ARL_0 (in gray) and ARL_1 values for disturbances in the MA coefficient θ_1

θ_1	n	I											
		10				30				100			
		$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\mathbf{e}}(\hat{\mu})$	$t_{\mathbf{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\mathbf{e}}(\hat{\mu})$	$t_{\mathbf{e}}(\hat{\sigma})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\mu})$	$\mathcal{T}^2_{\boldsymbol{\beta}}(\hat{\sigma})$	$t_{\mathbf{e}}(\hat{\mu})$	$t_{\mathbf{e}}(\hat{\sigma})$
0.0	100	1.04	0.05	189.46	187.56	1.03	0.02	318.45	171.10	1.02	0.02	341.67	138.61
	200	1.00	0.00	296.10	205.10	1.00	0.00	278.15	191.30	1.00	0.00	337.82	163.25
	500	1.00	0.00	220.48	197.75	1.00	0.00	243.65	186.73	1.00	0.00	416.67	131.76
	1000	1.00	0.00	201.59	189.57	1.00	0.00	238.89	184.90	1.00	0.00	322.92	203.17
0.3	100	7.33	7.74	170.40	184.38	5.84	3.00	205.65	180.53	4.74	1.68	253.01	185.03
	200	2.45	1.46	155.34	169.34	2.02	0.56	202.49	182.64	1.83	0.32	256.22	181.08
	500	1.07	0.08	185.49	189.92	1.04	0.03	190.98	165.04	1.03	0.02	210.29	158.95
	1000	1.00	0.00	148.41	172.60	1.00	0.00	211.30	174.23	1.00	0.00	268.31	195.02
0.4	100	55.40	78.49	129.41	146.02	32.90	27.50	147.97	161.64	27.84	16.54	181.80	156.61
	200	26.16	53.57	137.07	159.37	12.04	7.63	191.81	169.91	10.46	3.63	247.89	172.22
	500	4.26	2.38	178.77	187.13	3.42	1.38	183.42	173.98	3.02	0.81	224.55	173.42
	1000	1.61	0.47	149.25	168.66	1.50	0.35	159.74	163.63	1.37	0.15	253.06	173.74
0.5	100	169.63	153.31	112.62	152.65	127.89	119.85	155.47	166.63	126.76	108.87	216.21	171.22
	200	194.15	161.04	116.60	162.08	151.55	127.01	141.08	146.36	132.02	98.01	176.38	152.99
	500	197.73	162.23	115.91	150.33	171.52	142.86	156.87	157.14	154.73	124.42	204.45	163.41
	1000	177.92	167.45	133.91	166.74	167.25	146.04	146.48	158.55	136.12	115.48	200.83	156.53
0.6	100	51.96	61.10	114.11	137.23	39.03	54.42	93.17	111.18	25.69	11.13	161.06	152.71
	200	23.37	21.95	102.01	139.96	16.87	12.82	116.49	132.98	12.19	5.91	148.99	135.44
	500	4.76	2.88	99.08	116.60	3.69	1.98	113.42	119.19	3.18	1.02	138.46	136.73
	1000	1.49	0.36	98.71	124.72	1.44	0.30	109.10	119.26	1.33	0.15	165.97	152.86
0.8	100	2.35	1.74	70.42	116.22	1.86	0.75	83.91	121.46	1.53	0.34	88.00	122.71
	200	1.03	0.05	74.70	121.79	1.02	0.02	86.64	125.33	1.01	0.01	78.42	99.25
	500	1.00	0.00	74.36	125.29	1.00	0.00	73.80	100.02	1.00	0.00	89.07	115.21
	1000	1.00	0.00	75.09	108.74	1.00	0.00	83.58	122.84	1.00	0.00	90.60	106.09

Referências

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

[Box et al., 1970] Box, G. E., Jenkins, G. M., Reinsel, G., et al. (1970). Forecasting and control. *Time Series Analysis*, 3:75.

[Camacho et al., 2009] Camacho, J., Picó, J., and Ferrer, A. (2009). 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):59–68.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables with principal components. *J. Educ. Psy.*, 24:498–520.

[Hotelling, 1947] Hotelling, H. (1947). Multivariate quality control. *Techniques of statistical analysis.*

[Montgomery, 2007] Montgomery, D. C. (2007). *Introduction to statistical quality control.* John Wiley & Sons.

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

[Pearson, 1901] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

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

8 Conclusão

Este trabalho apresentou uma nova abordagem para lidar com processos em bateladas por meio de um conjunto de cartas de controle baseados no modelo ARMA. Foi construído um grupo de cartas com base nas estimativas dos coeficientes das bateladas históricas sob controle, que juntamente com as estatísticas de Hotelling e t-Stutent modificadas podem facilmente acomodar essas estimativas e uma regra de decisão foi feita para monitorar novas amostras.

Por meio de simulação, observamos que a carta T^2_β supera o concorrente tradicional com base nos resíduos para detectar mudanças de qualquer nível na dinâmica do processo, já as cartas t_β individuais são poderosas para identificar a fonte das perturbações. Além disso, a aplicabilidade da nossa abordagem foi ilustrada através de dados reais em que o bom desempenho é claramente notado.

Podemos elencar alguns possíveis trabalhos futuros, testar outros modelos de séries temporais, pensar essa situação com dados multivariados, bem como explorar e implementar aplicações reais.