Congresso de Métodos Numéricos em Engenharia 2015
Lisboa, 29 de Junho a 2 de Julho, 2015

COMBINAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS COM TÉCNICA NÃO PARAMÉTRICA PARA ESTABELECER VALORES LIMITE EM GRÁFICOS DE CONTROLE APLICADOS EM DADOS DE INSTRUMENTAÇÃO DE BARRAGEM

Emerson Lazzarotto[1]*, Liliana M. Gramani[2], Anselmo C. Neto[2], Luiz A. T. Junior[3] e Edgar M. C. Franco[1]

1: Centro de Engenharias e Ciências Exatas -Unioeste – Univ. Estadual do Oeste do Paraná
Dinter PPGMNE/UFPR – Unioeste - FPTI/BR, 85.870-650, Foz do Iguaçu – PR - Brasil
e-mail: {emerson.lazzarotto@gmail.com, emfra1@gmail.com}

2: Universidade Federal do Paraná – Setor de Ciências Exatas, Depto. Matemática e Estatística
PPGMNE - Centro Politécnico, Curitiba – PR - e-mail: {lgramani@gmail.com, anselmo@ufpr.br}

3: UNILA – Universidade Federal de Integração da América Latina
85.867-970, Foz do Iguaçu – PR - e-mail: luiz.a.t.junior@gmail.com

Palavras-chave: componentes principais, gráficos de controle multivariados, modelagem ARIMA-Garch, *kernel density estimation*, segurança de barragens.

Resumo. *O monitoramento de uma barragem requer muitos instrumentos para avaliar seu comportamento. Gráficos de controle multivariados são ferramentas estatísticas amplamente utilizadas para monitorar diversas variáveis de interesse. É necessário estabelecer valores estatísticos de controle com base no período operacional para avaliar a segurança de uma barragem. Assim, é desejável buscar a redução do número de variáveis, a descoberta de associações e o estabelecimento de valores de controle para parâmetros das novas variáveis do sistema reduzido. Gráficos de controle para algum parâmetro de interesse, como a média, requerem que os dados sigam a distribuição normal de probabilidades, o que geralmente não ocorre com dados de instrumentação. Uma alternativa para este tipo de problema nos dados é o uso dos resíduos de modelos ARIMA-Garch e de técnicas não paramétricas para estimar intervalos de controle. Este trabalho tem o objetivo de combinar o uso da técnica estatística multivariada de análise de componentes principais e de gráficos de controle em que os limites de controle para a média de um conjunto de resíduos de componentes principais são estabelecidos com base em critérios não paramétricos via kernel density estimation(kde) que independe da normalidade. Um estudo de caso realizado com dados de leitura da instrumentação da barragem da usina de Itaipu mostrou que o limite de controle estabelecido com base em critérios não paramétricos apresenta menor taxa de falsos alarmes estatísticos comparado com o método paramétrico e que a análise de componentes principais colabora na identificação de instrumentos responsáveis pela maior parte da variabilidade e que isto corrobora com sua posição relativa na barragem.*

1. INTRODUÇÃO

Garantir a preservação da segurança de uma barragem é uma tarefa fundamental em seu período operacional com vistas a fornecer alertas de emergência, minimizar a perda material e financeira e evitar vítimas humanas. A técnica para avaliar a segurança de uma barragem é a análise de sua instrumentação. Grandes barragens podem dispor de centenas ou mesmo milhares de instrumentos responsáveis pela avaliação de várias características de interesse da engenharia, como deslocamentos, subpressões e fornecem extensas bases de dados.

Devido às características geotécnicas específicas de cada obra de barragem, não se pode ter critérios de avaliação universais de barragens, ou seja, cada barragem tem suas especificidades. Além de instrumentar a barragem é preciso interpretar os dados de leituras dos instrumentos para avaliar seu comportamento e estabelecer, de alguma forma, um domínio de valores estatisticamente aceitáveis para as leituras de um dado instrumento em que se espera, com dada probabilidade, que a barragem esteja em condições normais de operação. Uma maneira de executar esta tarefa é utilizar ferramentas estatísticas aplicadas em dados de um período em que se considera a barragem sob controle para análise de seu comportamento futuro.

Em geral, mesmo instrumentos que avaliam aspectos diferentes, quando estão localizados próximos, apresentam correlação em suas leituras e estas são afetadas, sobretudo, por mudanças na temperatura e nível do reservatório. [1] Assim, pode ser útil o emprego de técnicas de controle estatístico de qualidade multivariada, que levem em conta a análise de um conjunto de instrumentos para estabelecer valores de controle para este conjunto, em vez da análise individualizada. Isto pode fornecer respostas mais globais e fidedignas sobre o comportamento da barragem e reduzir o trabalho.

Uma ferramenta de controle estatístico de qualidade bastante popular são os gráficos de controle que podem ser univariados ou multivariados. Sua concepção consiste de um valor central de uma característica de qualidade de interesse (como a média, por exemplo), de uma região (ou intervalo) de controle estatístico em que se estabelecem limites (superior/inferior) de controle e de uma probabilidade associada ao teste de hipótese que um conjunto de dados está sob controle. Para maiores detalhes se sugere a leitura de [2], [3].

Gráficos de controle tradicionais para um parâmetro de interesse foram concebidos sob a hipótese de que os dados têm distribuição normal de probabilidades e são independentes e identicamente distribuídos. Entretanto, tais hipóteses podem não serem verdadeiras quando se trata com dados de leituras de instrumentos. Não levar em conta tais hipóteses pode desacreditar o sistema de controle estatístico ao produzir excessivo número de alarmes falsos e também de falsos verdadeiros.

Para tratar de dados multivariados e superar estes problemas, vária abordagem tem sido propostas. [4] criou um modelo de monitoramento multivariado de dados de barragens extraindo componentes principais da matriz de dados e as interpretando como o efeito ambiental sobre os dados. O resíduo entre o vetor de dados e das componentes retidas leva a análise da estatística da norma quadrática e se obtém limites de controle na função densidade de probabilidades que é estimada pelo método *kernel density estimation (kde).*

Gráficos de controle não paramétricos das componentes principais de um sistema

multivariado em que não se requer a hipótese de normalidade foram desenvolvidos em [5]. Simulações da estatística T^2de Hotteling mostraram que o gráfico proposto tem melhor desempenho na ausência da normalidade.

Um estudo de caso industrial foi feito em [6], onde se combinou o uso de análise de componentes principais para definição dos limites de controle com um *kernel* de estimativa da densidade de probabilidade multivariada, seus resultados diminuíram o número de falsos alarmes, permitiram a identificação de comportamento anormal com antecedência e redução da variabilidade dos dados. Foram avaliadas as estatísticas SPE e T^2.

Para dados de um processo multivariado que mostram evidências de ausência da normalidade multivariada, o limite superior de controle do gráfico T^2 estabelecido com base num percentil da distribuição F, na fase II, pode não ser muito acurado. [7] usa a técnica *kde* para a distribuição da estatística T^2 e do limite superior de controle quando os dados não são normais multivariados.

Gráficos de controle para observações individuais de gráficos de médias móveis foram comparadas com gráficos não paramétricos baseados em quantis de distribuição *bootstrap* e *kernel estimators* em diversos cenários em estudos de simulação com dados de diversas distribuições paramétricas na fase II do gráfico de controle. [8]

O diagnóstico de valores singulares foi proposto no monitoramento de segurança de barragem, com um estudo de caso numa usina hidrelétrica na China, via análise de componentes principais e gráfico de controle multivariado T^2. [9] Foi aplicado numa usina chinesa um modelo que extrai as componentes principais dos dados da instrumentação e estabelecer um modelo hidrostático sazonal no tempo entre nível do reservatório, temperatura e efeitos do tempo e as componentes principais. Após uma regressão entre as variáveis podem ser feitas previsões das componentes principais e avaliar por meio de gráficos de controle T^2 o comportamento dos instrumentos. [10]

Neste trabalho se busca estabelecer limites de controle para o gráfico multivariado dos resíduos de componentes principais (PCA) retidas de um processo multivariado. Os resíduos são obtidos após uma modelagem ARIMA-Garch das componentes. Os limites de controle são obtidos através da técnica não paramétrica KDE utilizando dois núcleos e duas variações do parâmetro h. Uma comparação com os limites obtidos com a hipótese de normalidade multivariada (MVN) para dois tipos de estimativas da matriz de covariância dos dados é realizada em dados de monitoramento da barragem da usina de Itaipu entre o Brasil e o Paraguai. Os resultados mostram que a combinação de PCA-ARIMA-Garch e KDE facilita a redução da massa de dados, a obtenção de variáveis representativas e a redução de falsos alarmes para estabelecer limites de controle.

Este artigo está assim estruturado: na seção 2 é feito uma breve explanação sobre gráficos de controle, análise de componentes principais, modelagem ARIMA-Garch e KDE; na seção 3 são descritos os dados e métodos empregados; na seção 4 são apresentados os resultados obtidos e na seção 5 algumas conclusões e considerações.

2. REREFENCIAL TEÓRICO

2.1. Gráficos de controle

De acordo com [2], uma das principais técnicas de controle estatístico do processo é o gráfico (ou carta) de controle. Um típico gráfico de controle consiste em plotar a média da medida de uma característica de qualidade em função do tempo. O gráfico possui uma linha central que representa o valor alvo caso não houvesse variabilidade e de duas outras linhas, o limite inferior e superior de controle que são determinados com base estatística. Os gráficos de controle são elementos visuais para o monitoramento da conformidade de características de produtos ou processos. [11]

Existe a ocorrência de causas comuns de variação, que são inerentes ao processo e de causas especiais de variação, devido a mudanças reais. Quando somente as comuns estiverem presentes o processo é estável ou sob controle estatístico e quando há causas especiais de variação o processo é instável ou fora de controle estatístico. [12]

Os gráficos de controle podem ser úteis para verificar se dados passados se originavam de um processo sob controle ou para indicar se amostras futuras deste processo estão sob controle estatístico. [2]

Quando se estabelece um teste de hipótese sobre uma afirmação H_0: um processo está sob controle, tem-se associado a ele dois tipos de erros, a saber,

$$\begin{aligned} \alpha &= P(erro\ tipo\ I) = P(rejeitar\ H_0 | H_0 \text{ é } verdadeira) \\ &= P(rejeitar\ o\ estado\ de\ controle | processo\ está\ sob\ controle) \\ &= P(falso\ alarme) \end{aligned}$$

α é chamado de nível de significância do teste e

$$\begin{aligned} \beta &= P(erro\ tipo\ II) = P(aceitar\ H_0 | H_0 \text{ é } falsa) \\ &= P(aceitar\ o\ estado\ de\ controle | processo\ está\ fora\ de\ controle) \\ &= P(falso\ positivo) \end{aligned}$$

Quando se aumenta o intervalo ao redor da linha central de um gráfico de controle é diminuído o risco de erro tipo I e é aumentado o risco de erro tipo II, por outro lado quando se diminui o intervalo ao redor da linha central de um gráfico de controle é aumentado o risco de erro tipo I e é diminuído o risco de erro tipo II. Dependendo da necessidade que se tem, pode-se fixar o valor do múltiplo do desvio padrão e calcular o valor de α ou vice-versa. [2]

Cometer um erro tipo I significa apontar um falso alarme, ou seja, rejeitar que o processo está sob controle num dado instante quando, na verdade, está sob controle. Um importante instrumento para medir o desempenho (reduzir a taxa de falsos alarmes e aumentar a capacidade de medir mudanças da característica de interesse) de um gráfico de controle é o comprimento médio da seqüência (CMS) que é número médio de observações que se espera até que uma observação indique uma condição de fora de controle.

Quando o processo está sob controle é desejável aumentar o tempo necessário para que o gráfico de controle emita um alarme falso para reduzir a taxa de falsos alarmes (aumentar o CMS). No gráfico de controle de um processo não autocorrelacionado o CMS é calculado por $CMS = 1/p$, onde $p = P$(uma observação apareça fora de controle). Por exemplo, o gráfico de controle da média amostral com limites de $\pm 3\sigma$ onde $p = 0{,}0027, CMS = 370$, significa

que, mesmo que o processo esteja sob controle, *em média,* a cada 370 observações espera-se que uma observação esteja fora do intervalo $(\bar{x} - 3\sigma, \bar{x} + 3\sigma)$. Os limites de controle dependerão da hipótese de normalidade, do nível de significância α, do tamanho n da amostra e da quantidade m de amostras. [2]

O planejamento e a análise de um gráfico de controle são constituídos de duas fases. Na fase I, chamada retrospectiva, quando não é possível especificar valores padrão ou de referência para a média e o desvio padrão do processo (maioria das vezes), são estabelecidos os limites de controle com base em amostras. [13] Na fase II, chamada de fase prospectiva, o monitoramento do processo continua baseado em novos dados de entrada.

O conceito básico do controle estatístico de processos baseia-se na comparação do que está acontecendo hoje com o que aconteceu previamente. Uma componente importante de um controle de processo é a obtenção de uma base de dados do processo sob controle que serve de referência ou calibração, chamada de fase I. Isto quer dizer que o processo está operando próximo de um valor alvo aceitável com alguma variação natural e sem causas de preocupação. [13]

O uso de gráficos de controle em certos dados ambientais é insatisfatório e fornece muitos falsos alarmes devido ao fato dos dados serem correlacionados. Este problema pode ser resolvido pelo ajuste de um modelo de série temporal usando os resíduos do modelo no monitoramento, onde o resíduo é a diferença entre o valor observado e o valor previsto. [3]

A alta quantidade de características monitoradas e o recebimento *on line* de dados de monitoramento fez crescer no meio industrial e acadêmico, nos últimos 30 anos, o interesse pelo controle estatístico de processos multivariados. Na prática quase sempre há necessidade de monitorar o controle de diversas variáveis e, embora múltiplos gráficos de controle univariados possam ser aplicados, isto pode conduzir a interpretações enganosas sobre o estado de um processo, sobretudo quando existir correlação entre as variáveis, as variáveis devem ser examinadas conjuntamente e não separadamente. [14]

Se p variáveis representam processos que estejam sendo controladas, assumindo que $\boldsymbol{X}' = [X_1, X_2, \dots, X_p] \sim N_p(\mu, \Sigma)$, ou seja, que $\boldsymbol{X}'$ tenha distribuição normal multivariada e que se deseja controlar a média μ, as covariâncias entre as variáveis X_i e as variâncias $V(X_i)$. A alteração de ao menos uma das médias ou das covariâncias (variâncias) significa que o processo está fora de controle. Neste caso, [2] e [15] relatam que a aplicação de gráficos de controle univariados pode conduzir a interpretações equivocadas e enganosas e que os métodos multivariados são uma boa alternativa. Quando as variáveis são correlacionadas aumenta a probabilidade de emissão de falsos alarmes ou, pior ainda, de não receber um sinal de alerta quando o processo multivariado esteja fora de controle.

A estatística T^2 pode ser considerada uma generalização da estatística t que possui distribuição t-de Student onde $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$. Note que t é a distância entre a média amostral e da população, ponderada pelo desvio padrão amostral. Quando se deseja testar a hipótese $\mu = \mu_0$ tem-se que

$$t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n} = n(\bar{x} - \mu_0)[s^2]^{-1}(\bar{x} - \mu_0).$$

Logo, no caso p dimensional, se S^{-1} é a inversa da matriz de covariância, segue que

$$T^2 = n(\bar{x} - \mu_0)'[S]^{-1}(\bar{x} - \mu_0).$$

No caso de m amostras, de um processo sob controle, extraídas para avaliação serem de tamanho n, se p variáveis estiverem sendo avaliadas, de acordo com [2], como na prática quase sempre μ e Σ são desconhecidas e são estimadas, respectivamente, pelos estimadores não viesados $\bar{\bar{x}}' = (\bar{\bar{x}}_{\mathbf{1}}, \dots, \bar{\bar{x}}_{\boldsymbol{p}})$ e a matriz simétrica positiva definida S de covariância, então a estatística torna-se $T^2 = n(\bar{x} - \bar{\bar{x}})'[S]^{-1}(\bar{x} - \bar{\bar{x}})$. Esta estatística é chamada de T^2 de Hotteling ou somente T^2 e tem o objetivo de medir a distância de cada observação do centroide dos dados, em uma escala padronizada que compensa a diferença em unidades de medida de variabilidade. Quando o número de observações de uma amostra é $n = 1$, a estatística T^2 de Hotelling conforme [2], [15] e [3] fica

$$T^2 = (x - \bar{x})'[S]^{-1}(x - \bar{x})$$

e o limite superior de controle da fase II torna-se

$$LSC = \frac{p(m+1)(m-1)}{m(m-p)} F_{\alpha,p,m-p}. \tag{1}$$

Seguindo a indicação de [16], o limite superior de controle (LSC) da fase I, no caso de observações individuais é recomendável calcular se baseando na equação

$$LSC = \frac{(m-1)^2}{m} \beta_{\alpha,\frac{p}{2},\frac{m-p-1}{2}} \tag{2}$$

e o elipsóide p dimensional de $100(1-\alpha)\%$ de previsão de uma futura observação é dado por todos os vetores x satisfazendo $(x - \bar{x})'[S]^{-1}(x - \bar{x}) \leq LSC$ onde m é a quantidade de amostras.

Uma questão importante no caso de tratamento de processo com observações individuais é a forma de estimar a matriz de covariância. O estimador usual é aquele obtido da combinação das m observações

$$S_1 = \frac{1}{m-1}(x_i - \bar{x})(x_i - \bar{x})'. \tag{3}$$

É sugerido por [17] o uso do estimador da matriz de covariância a partir de pares de diferenças sucessivas, dado por

$$\boldsymbol{v}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i \qquad i = 1, \dots, m-1.$$

Os vetores $\boldsymbol{v}_i$ podem ser arranjados em uma matriz

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1{}' \\ \vdots \\ \boldsymbol{v}_{m-1}{}' \end{bmatrix}$$

De modo que a matriz

$$S_2 = \frac{1}{2(m-1)} \boldsymbol{V}'\boldsymbol{V} \tag{4}$$

é um estimador mais ‘local’ da matriz de covariância, no sentido de capturar somente variabilidade de curto prazo.

A estatística T^2 tem a vantagem de poder ser decomposta em componentes que refletem a contribuição individual de cada variável para o valor de T^2. Segundo [18],

$$d_j = T^2 - T^2_{(j)} \tag{5}$$

onde $T^2_{(j)}$ é o valor da estatística usando todas as variáveis com exceção de j, é um indicador da contribuição relativa da j-ésima variável na estatística global. Um valor alto de d_j indica forte contribuição desta variável numa observação eventualmente fora de controle e pode conduzir a análise desta variável em particular.

2.2. Análise de Componentes Principais

Técnicas de redução de dimensionalidade de dados são baseadas no princípio da criação de conjuntos de variáveis latentes, chamadas de componentes principais, que capturam a variação significativa que está escondida nos dados. Os escores das variáveis fazem parte destes conjuntos de variáveis latentes. Para o monitoramento do processo, a variação que os conjuntos de variáveis latentes extraem das variáveis de processo é de fundamental importância para a avaliação da qualidade do produto, da segurança do processo e, mais geralmente, se o processo está em controle estatístico. [19]

Um método que consiga extrair características nos dados pode ser útil no estudo de segurança de barragem. Como as leituras de instrumentação são resultado da combinação de diversos fatores, os métodos de análise multivariada dos dados podem apresentar as seguintes vantagens: 1) mais rentáveis ao reduzir o número de análises individuais, 2) maior capacidade de explicar e separar a variabilidade devida a uma causa atribuível da variabilidade aleatória dado que as componentes principais são, por definição, não correlacionadas e 3) identificar padrões de comportamento. [20,21]

Análise de componentes principais (PCA) é uma técnica analítica multivariada de dados na qual um número de variáveis relacionadas são transformadas em um conjunto em que se espera um número menor de variáveis não correlacionadas que são combinações lineares das variáveis originais. [22] [15] Sua popularidade provavelmente deriva de sua habilidade, em muitas ocasiões, de representar uma situação multivariada em um espaço de dimensão muito mais reduzida. [22]

Dado um vetor de variáveis aleatórias $X' = (X_1, X_2, \dots, X_p)$ com matriz de covariância conhecida Σ e autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, ao se considerar combinações lineares do tipo$Z_i = \boldsymbol{a_i'}X = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, i = 1, \dots, p$, usando propriedades da variância se mostra que $Var(Z_i) = \boldsymbol{a_i'}\Sigma\boldsymbol{a_i}$ e $Cov(Z_i, Z_k) = \boldsymbol{a_i}'\Sigma\boldsymbol{a_k}$.

Para $i, k = 1, \dots, p$, para que as novas variáveis (componentes principais – CPs) Z_i tenham variância máxima possível e sejam não correlacionadas, restringindo que os vetores $\boldsymbol{a_i}$ tenham norma unitária, se demonstra em [15] que a i-ésima componente principal $i = 1, \dots, p$, é dada por

$$Z_i = \boldsymbol{e_1}'X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p$$
$$Var(Z_i) = \boldsymbol{e_i'}\Sigma\boldsymbol{e_i} = \lambda_i$$

onde $(\lambda_i, \boldsymbol{e}_i)$ representa o i-ésimo par ordenado de autovalor (em ordem decrescente de valor) e autovetor da matriz de covariância. Além disso, a proporção total de variância explicada pela k-ésima componente principal é

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

e normalmente "poucas" k componentes principais conseguem explicar grande parte da variabilidade, de modo que as p variáveis originais podem ser subsituídas pelas k componentes principais com pouca perda de informação.

Existe uma série de critérios para a escolha do número de componentes principais a ser tomada. Estes critérios incluem: a) testes de significância para a igualdade de raízes, b) reter uma quantidade de CPs suficientes para explicar uma proporção determinada da variância total e c) parar quando as variâncias residuais forem iguais a alguma quantidade determinada, usualmente a variabilidade inerente do sistema em estudo. [23]

Para [15], onde há necessidade de se monitorar a qualidade de um grande número de variáveis, o uso de CPs combinado com gráficos de controle multivariado é uma alternativa plausível e viável. Em um processo que se encontra estável no tempo, no qual as características avaliadas são influenciadas apenas por causas comuns de variação , os valores de algumas "poucas" componentes principais também deverão estar estáveis. Reciprocamente, se as componentes principais apresentam estabilidade, a variação aleatória das variáveis originais também deverá estar estável.

2.3. Modelagem ARIMA-GARCH

A modelagem de séries temporais sempre atraiu a atenção de diversos pesquisadores devido a sua importância na aplicação em quase todas as áreas do conhecimento. O objetivo é se ajustar um modelo adequado a um conjunto de dados disponíveis, bem como, que seja capaz de ter um desempenho o mais exato e preciso possível na previsão de valores da série.

De acordo com [24], são objetivos das séries temporais: investigar seu mecanismo gerador, fazer previsões de valores futuros, descrever o seu comportamento, procurar periodicidades relevantes. Uma série temporal é estacionária de segunda ordem, se flutua estocasticamente ao redor de uma média constante e possui autocovariancia constante.

A abordagem de Box e Jenkins [25] é uma metodologia bastante utilizada na análise de modelos paramétricos e consiste do ajuste de modelos auto-regressivos integrados de médias móveis ($ARIMA(p,d,q)$) que são modelos lineares aplicáveis em séries estacionárias de 2ª ordem. Sua construção se baseia num ciclo iterativo cujos estágios estão representados na Figura 1.

Denotando o operador de translação ao passado por B, pode-se escrever que $BZ_t = Z_{t-1}$, $B^m Z_t = Z_{t-m}$ e o operador diferença Δ, de modo que $\Delta Z_t = Z_t - Z_{t-1} = (1-B)Z_t$, os modelos lineares estacionários $ARMA(p,q)$ tem a forma

$$Z_t = \mu + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}.$$

Onde μ, ϕ_i e θ_i são parâmetros a serem determinados via mínimos quadrados e os resíduos a_t são não correlacionadas . A identificação de modelos $ARIMA$ para ajustar um conjunto de dados tem o objetivo de determinar os valores iniciais dos parâmetros p, d e q e, em seguida, estimar os parâmetros a serem usados na etapa de estimativas. A identificação é

feita, sobretudo, com base na análise de perfil dos gráficos das funções de autocorrelações simples e autocorrelações parciais estimadas, as quais se espera representarem adequadamente seus valores teóricos que são desconhecidos. [24]
As etapas da identificação são:analisar se há a necessidade de aplicar alguma transformação (logarítmica, Box-Cox,...) nos dados da série original com vistas a estabilizar sua variância, aplicar diferenças na série para obtenção de estacionariedade e redução a busca de um modelo $ARMA(p,q)$; fazer a identificação dos valores de p e q com base nas autocorrelações e autocorrelações parciais.

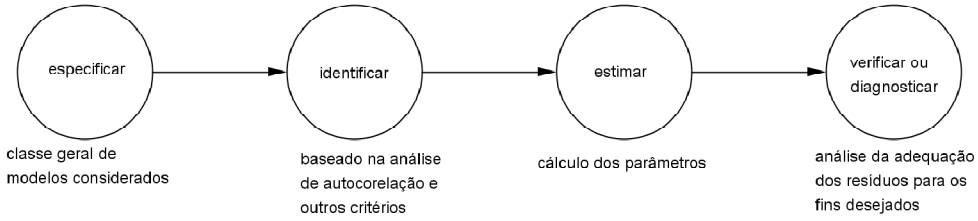

Figura 1 - Estágios do ciclo iterativo da construção de um modelo de séries temporais

Quando se tem séries temporais cuja variância condicional evolui no tempo, modelos lineares como $ARIMA$ podem não mais ser adequados para descrever satisfatoriamente o seu comportamento estocástico. Modelos não-lineares podem ser mais apropriados neste caso. [24] Por exemplo, modelos $ARCH$, que são não-lineares no que se refere à variância. A variância condicional também é chamada de volatilidade. Os modelos auto-regressivos com heteroscedasticidade condicional ($ARCH$) foram introduzidos por [26] com o objetivo de estimar a média e a variância da inflação.
A partir daí, surgiram variantes do modelo original de heteroscedasticidade, em particular os modelos $GARCH$($ARCH$ generalizado) proposto por [27], cujo modelo mais simples é o $Garch(1,1)$, expresso como $\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 \sigma_{t-1}^2 \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 \sigma_{t-1}^2$, em que a variância condicional de u no tempo t não depende somente do termo do erro ao quadrado no período anterior (como ocorre no $ARCH(1)$), mas também da variância condicional no período anterior. Este modelo pode ser generalizado ao caso $Garch(p,q)$. [28]

2.4. Kernel Density Estimation (KDE)

Quando uma amostra é extraída de uma distribuição contínua, deseja-se estimar a distribuição da população do qual aquela amostra é extraída. Caso se disponha de uma estimativa da densidade de uma população contínua, é possível determinar estimativas das estatísticas da população tais como a média, moda, domínio, quantis e da simetria da distribuição. [29] Isto é particularmente útil quando a distribuição é desconhecida e requer cálculos computacionalmente intensivos.
Um núcleo estimador da densidade (KDE) é uma maneira não paramétrica de estimar a função densidade de probabilidade de uma variável aleatória. A técnica KDE oferece uma maneira não paramétrica de estimar a função densidade de probabilidade sem a especificação de um modelo paramétrico (quando a família de funções de um modelo pode ser especificada por um número finito de parâmetros). [30]

Uma núcleo-função é uma função K tal que $K(x) \geq 0, x \in \mathbb{R}, K(x) = K(-x)$ e $\int_{-\infty}^{\infty} K(x)dx = 1$. Alguns exemplos (dentre vários existentes) de núcleo-funções são o núcleo de Epanechnikov

$$K_e(x) = \begin{cases} \frac{3(1-x^2)}{4}, & -1 \leq x < 1 \\ 0, & caso\ contrário \end{cases} \quad (6)$$

e o núcleo oriundo da distribuição normal

$$K_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, x \in \mathbb{R}. \quad (7)$$

A escolha de um modelo paramétrico depende da situação, seja por razões científicas ou de experiências anteriores. Isto pode ser uma desvantagem dos modelos paramétricos em restringir a algumas famílias paramétricas que podem não ser adequadas ao problema em que se está tratando. Os modelos não paramétricos dispensam a restrição a certas famílias paramétricas e buscam "deixar os dados falarem por si próprios". [30]

Um problema de regressão muito comum é a partir dos dados de amostra aleatória estimar a função densidade de probabilidade (fdp). No caso paramétrico, tal estimativa assume que a função densidade pertence a alguma família paramétrica, como a normal, e estima os parâmetros usando, por exemplo, estimadores de máxima verossimilhança. No caso não paramétrico a estimativa não assume nenhuma forma prévia da fdp. O histograma é um exemplo de estimativa da fdp não paramétrica em que a reta real é dividida em intervalos de tamanhos iguais e a fdp é uma função degrau com as alturas sendo da amostra contida naquele intervalo dividido pela largura do intervalo. [30]

A escolha da forma da núcleo função não é particularmente importante. Contudo, a escolha do valor do tamanho do intervalo h é muito importante. Quando h é grande, diminui a variância e aumenta o viés, já quando h é pequeno, aumenta a variância e diminui o viés, o h ótimo busca o melhor 'custo-benefício' entre variância e viés.

3. DADOS E MÉTODOS

Foram utilizados neste trabalho dados provenientes da leitura de piezômetros localizados no trecho E da barragem da usina hidrelétrica de Itaipu Binacional, localizada no Rio Paraná, entre o Brasil e o Paraguai, conforme a Figura 2. Itaipu foi até 2013 a maior geradora de energia elétrica do mundo e sua geração atende cerca de 17% da energia elétrica anualmente consumida no Brasil e 75% do Paraguai.

Os piezômetros são instrumentos responsáveis por medir subpressões no contato concreto-rocha e em níveis mais permeáveis do maciço basáltico da fundação provenientes de infiltrações. Estão disponíveis 319 observações de sete piezômetros (designados daqui em diante por $p_1, p_2, \ldots, p_7$) locazidos no trecho E6 no período de 2001 a 2013 de periodicidade aproximadamente quinzenal. Os sete instrumentos possuem de média a elevada correlação. Foram separadas 300 observações para a fase I de ajustamento de modelos e 19 para a fase II de teste(previsão). Uma visão da seção do trecho E6 e a

localização dos piezômetros é dada na Figura 3.
É usual calcular as componentes principais baseadas em variáveis originais padronizadas, isto é, com média zero e variância um, caso em que a matriz de covariância está na forma de matriz de correlação. A razão para tal procedimento é que as variáveis originais podem possuir escalas de domínio e magnitude bastante distintas dando falsa interpretação da sua real variabilidade, o que é evitado pela padronização. [2]
Com o objetivo de 'reduzir' os dados e descobrir novos conhecimentos, foram obtidas as CPs do conjunto de dados da fase I. As componentes retidas foram então analisadas quanto à existência de autocorrelação e de parte sistemática a ser modelada via modelos ARIMA-Garch.
A modelagem ARIMA-Garch das componentes principais é avaliada através dos correlogramas dos resíduos e dos resíduos quadrados, uma 'boa' modelagem deve produzir correlogramas em que não exista correlação significativa, o teste estatístico de Durbin Watson deve estar 'próximo' de 2 e os termos ARMA e Garch devem ser significativos (valor $p < 0{,}05$).

Figura 2 - Barragem da usina hidrelétrica de Itaipu Binacional

Obtidas as séries de resíduos das componentes principais, denominadas res_pc1, res_pc2, res_pc3 e res_pc4, os resíduos do modelo ajustado são avaliados quanto à ausência de correlação, estacionariedade (teste ADF – Dickey Fuller) e independência (teste BDS) e de normalidade (teste Jarque-Bera).
Os resíduos das componentes principais são usados na sequencia para construção do gráfico de controle multivariado T^2. Para a fase I, foi construído o gráfico T^2 com a presunção de normalidade multivariada (MVN) dos dados usando a matriz de covariância usual e de diferenças sucessivas. (equações (3) e (4))
Os valores da estatística T^2 na fase I, para ambas as matrizes de covariância foram usados para a construção de uma densidade de probabilidades empírica via *kde* usando os núcleos de Epanechnikov e da normal, usando 100, 200 ou 300 pontos de malha e dois tamanhos

para o comprimento do parâmetro h da estimativa, um 'ótimo' fornecido pelo *software Eviews* e outro sendo $h/2$, para avaliar diferentes cenários.
Limites superiores de controle para todos estes cenários ao nível de erro tipo I, $\alpha = 1$, $\alpha = 2$ e $\alpha = 5$ foram determinados para os dados da fase I e da fase II. Foi avaliada e comparada a quantidade de observações fora do limite de controle (FLC) para ambas as fases em todos os cenários. Um resumo das etapas do processo empregado é apresentado na Figura 4.

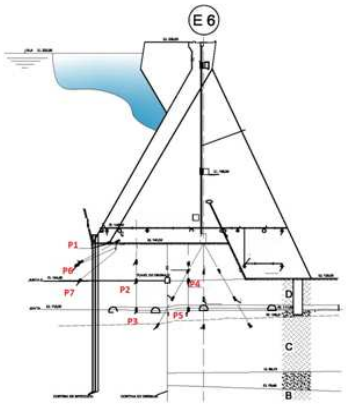

Figura 3 - Seção do bloco 6 do Trecho E de Itaipu com instrumentos p1 até p7 em destaque

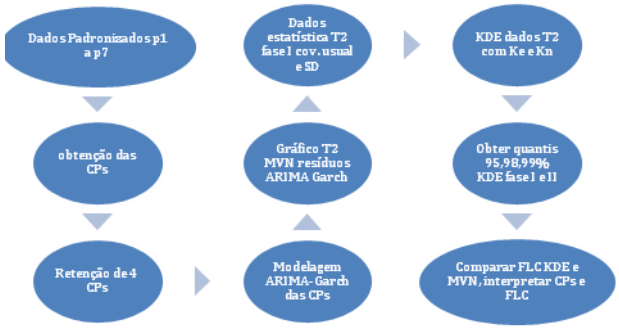

Figura 4 - Fluxograma do método empregado

4. RESULTADOS

Considerando o conjunto de dados da fase I, na Tabela 1, Tabela 2 e Tabela 3 são apresentados os autovalores, autovetores, correlações e o teste de normalidade univariada de Jarque-Bera para cada componente principal (aqui rotuladas como pc1 até pc7).

Comp.	Autovalor	Proporção	Prop. Acum.
1	3,5558	0,5080	0,5080
2	1,4512	0,2073	0,7153
3	1,0225	0,1461	0,8614
4	0,5017	0,0717	0,9331
5	0,2640	0,0377	0,9708
6	0,1359	0,0194	0,9902
7	0,0684	0,0098	1,0000

Variável	p_1	p_2	p_3	p_4	p_5	p_6	p_7
p_1	-0,1029	0,6934	-0,3506	0,4858	-0,0651	0,3719	0,0838
p_2	0,3485	0,4288	0,1238	-0,7007	-0,3265	0,2820	-0,0454
p_3	0,4480	-0,3430	0,1700	0,1886	0,1189	0,5824	0,5135
p_4	0,1457	0,4160	0,7757	0,2213	0,3095	-0,2425	0,0160
p_5	0,4587	-0,1106	0,0659	0,4221	-0,6996	-0,2262	-0,2324
p_6	0,4455	0,1774	-0,4145	-0,0886	0,1897	-0,5599	0,4906
p_7	0,4867	-0,0274	-0,2342	0,0477	0,5036	0,1390	-0,6574

Tabela 1 - Análise das componentes principais de p_1 a p_7 autovalores (esquerda) e autovetores (direita)

Correlações	p_1	p_2	p_3	p_4	p_5	p_6	p_7
p_1	1,000000						
p_2	0,108466	1,000000					
p_3	-0,493776	0,307388	1,000000				
p_4	0,123649	0,423968	0,171992	1,000000			
p_5	-0,200625	0,411950	0,789374	0,220192	1,000000		
p_6	0,113738	0,601896	0,520020	0,033913	0,626103	1,000000	
p_7	-0,115457	0,503588	0,756680	0,091113	0,705906	0,853915	1,000000

Tabela 2 - Análise das componentes principais de p_1 a p_7 - Correlações

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Jarque-Bera	12,75064	1,649270	6,661485	5,621453	2,534641	4,549800	80,43267
Probabilidade	0,001703	0,438395	0,035767	0,060161	0,281585	0,102807	0,000000

Tabela 3 - Análise das componentes principais de p_1 a p_7 - Teste de normalidade

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
Jarque-Bera	143,6852	15,61374	12,50464	328,0493	84,72870	16,55980	12,40401
Probabilidade	0,000000	0,000407	0,001926	0,000000	0,000000	0,000254	0,002025

Tabela 4 - Teste de normalidade Jarque-Bera aplicado aos dados originais padronizados (p_1 a p_7)

Observa-se que as CPs apresentam melhoria quanto à normalidade em relação aos dados dos instrumentos originais, ao nível de confiança de 95%, todos os testes de Jarque-Bera de normalidade (Tabela 6) rejeitam a hipótese sob os instrumentos (valor $p < 0{,}05$) e para as CPs 2,4,5 e 6 já se pode aceitar a hipótese de normalidade. Foram selecionadas 4 CPs tendo em vista que elas explicam mais de 90% da variabilidade, conforme Tabela 1.

A análise dos correlogramas das CPs mostrou que existe forte autocorrelação em todas elas e na Figura 5 é exibido o correlograma de pc1, o que indica que a série pode ser mais bem modelada via método ARIMA-Garch.

A Tabela 5 apresenta o resultado da avaliação da significância dos coeficientes ARIMA-Garch(1,1) da variável pc1 e a Figura 7 o correlograma dos resíduos e resíduos quadráticos pós modelagem ARIMA-Garch, note que não há correlações significativas (valor $p < 0{,}05$) o que indica que o modelo é adequado. Para as demais componentes pricipais foi executado procedimento semelhante de modelagem ARIMA-Garch e obtenção dos resíduos da modelagem que foram denominados de res_pc1, res_pc2, res_pc3 e res_pc4.

A Tabela 7 apresenta o resultado do teste de normalidade de Jarque-Bera para normalidade univariada. Nota-se que somente um dos resíduos das CPs apresenta distribuição normal ao nível de significância de 95% e como as distribuições marginais serem normais é condição necessária para a distribuição conjunta das 4 CPs serem normais multivariadas, podemos rejeitar esta hipótese.

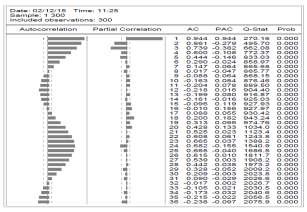

Figura 5 - Correlograma de pc1 (fac e facp)

Os resíduos da modelagem ARIMA-Garch devem ser entendidos como o resíduo da modelagem de um conjunto de instrumentos dos quais foram retidas componentes responsáveis por grande parte da variabilidade nos dados, que são não correlacionadas e das quais foi retirada uma variação sistemática (modelagem ARIMA-Garch) possivelmente oriunda de fatores ambientais (temperatura) e do nível do reservatório.
Sob estas considerações, este conjunto de resíduos deve ser uma indicação do controle estatístico do conjunto de instrumentos, ou seja, se estes estiverem sob controle do ponto de vista estatístico, a estrutura de variabilidade deste conjunto de instrumentos também deve ter se mantido, caso contrário os dados do conjunto de instrumento estariam estatisticamente fora de controle.
A etapa final é a construção do gráfico de controle da estatística T^2 para as fases I (300 dados) e II (19 dados) e avaliar se o conjunto de dados dos resíduos das componentes principais (res_pc1 a res_pc4) está ou não sob controle do ponto de vista estatístico. Os limites de controle para o gráfico T^2 quando se tem normalidade multivariada (que não é o caso!), conforme as equações (1) e (2) se baseiam numa distribuição F ou β.
Com os dados de referências da estatística T^2 seja ela calculada via matriz de covariancia usual ou de diferenças sucessivas equações (3) e (4) na fase I, foram construídas distribuições de probabilidade empíricas com *kernel density estimation* usando os núcleos de Epanichinikov e da distribuição normal (equações (6) e (7)) variando os parâmetros número de pontos na malha de 100, 200 ou 300 e tamanho do intervalo h (default do software Eviews) e $h/2$ para os resíduos das CPs retidas (1 a 4). A Figura 6 contém o histograma de dados de T^2 e duas *kernel* funções de estimativa da densidade.

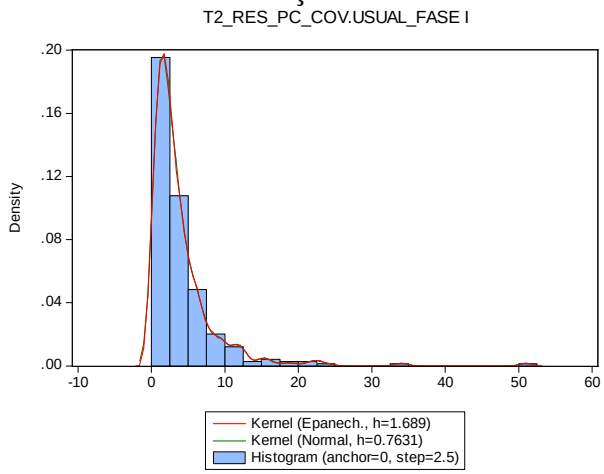

Figura 6 - Exemplo de kernel funções de estimativa da densidade de probabilidade dos valores de T2

Os resultados do número de observações fora do limite de controle (FLC) para cenários em que se estabelece a taxa α de erro tipo I (ou de falsos alarmes) em $1\%, 2\%$ e 5% do gráfico T^2 dos resíduos de pc1 a pc4 são apresentados na Tabela 8 (covariância usual) e Tabela 9 (covariância por diferenças sucessivas) juntamente com o respectivo número de FLC considerando normalidade multivariada (MVN) e o número médio esperado (CMS).

Variável	Coeficiente	Erro padrão	Estatística-z	Probabilidade
AR(1)	1,231	0,033136	3,715,668	0.0000
AR(3)	-0,32917	0,034245	-9,612,359	0.0000
MA(1)	-0,29899	0,056517	-5,290,262	0.0000
MA(4)	0,11745	0,057654	2,037,185	0.0416
MA(23)	0,34444	0,039552	8,708,724	0.0000
C	0,19516	0,014555	1,340,822	0.0000
RESID(-1)^2	-0,03204	0,002466	-1,299,497	0.0000
RESID(-2)^2	0,18046	0,065541	2,753,458	0.0059
R-quadrado	0,93491	Estatística Durbin Watson		1,9387

Tabela 5 - Resultados da modelagem ARIMA-Garch de pc1

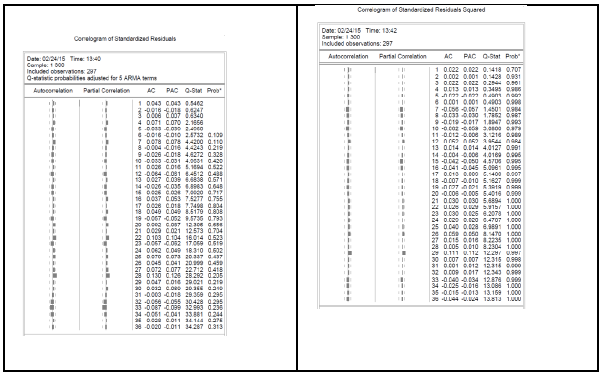

Figura 7- Correlograma dos resíduos e resíduos quadráticos de pc1

	PC1	PC2	PC3	PC4
Jarque-Bera	13.17534	0.861922	7.636870	4.903472
Probability	0.001377	0.649884	0.021962	0.086144

Tabela 6 - Teste de normalidade Jarque-Bera para as componentes retidas

Observa-se que os resíduos das PCs são mais fortemente desviados da normalidade e que os resíduos são dados mais 'refinados' o que provavelmente acarreta um maior número de FLC (fase I e II) quando se usa o gráfico T^2 comparativamente as PC originais (linhas 13,14 Tabela 8 e Tabela 9). Entretanto, ao usar a matriz de covariância estimada pela equação (4), observa-se que o número de FLC para o gráfico T^2 das CPs é completamente inviável, do que se conclui que, além do que já foi citado, é importante a modelagem ARIMA-Garch pois gera resíduos menos susceptíveis a forma com que se estima a matriz de covariância, com maior capacidade de construir um conjunto histórico de dados fidedigno e capaz de detectar mudanças na estrutura de variabilidade.

Quanto à análise de controle estatístico, observa-se que o gráfico T^2 dos resíduos de pc1 a pc4 apresenta número de FLC muito próximo do desejável do qual se pode concluir estatisticamente que os dados da fase I e da fase II estão sob controle, independente da matriz de covariância utilizada, quando se estima o limite superior de controle (LSC) via técnica não paramétrica *kde* e se usa uma malha de pontos maior (300) e que o tamanho do parâmetro h

provocou quase nenhuma alteração nos resultados. Este comportamento deu-se igualmente para as 3 taxas de falsos alarmes fixadas.

	RES_PC1	RES_PC2	RES_PC3	RES_PC4
Jarque-Bera	144.9864	3350.670	550.8130	3.491605
Probability	0.000000	0.000000	0.000000	0.174505

Tabela 7 - Teste de normalidade Jarque Bera para os resíduos do modelo ARIMA-garch das componentes retidas

		1%			2%			5%		
Covariância Usual		LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II
T2 Resíduos PC1 a PC4 - 297 obs.	KDE-E-100	18,49	6	7	15,14	10	12	11,22	18	20
	KDE-E-200	21,52	5	5	16,29	7	9	11,61	17	19
	KDE-E-300	22,14	4	4	16,84	7	9	11,73	16	18
	KDE-N-100	19,22	6	7	15,38	8	10	11	18	21
	KDE-N-200	21,84	5	5	16,39	7	9	11,49	18	20
	KDE-N-300	22,16	4	4	17,27	7	9	11,65	17	19
	KDE-E-100 h/2	18,81	6	7	15,06	10	12	11,31	18	20
	KDE-E-200 h/2	21,65	5	5	15,78	7	9	11,52	18	20
	KDE-E-300 h/2	22,08	4	4	17,15	7	9	11,68	16	18
	KDE-N-100 h/2	18,81	6	7	15,06	10	12	11,3	18	20
	KDE-N-200 h/2	21,69	5	5	15,89	7	9	11,66	17	19
	KDE-N-300 h/2	22,08	4	4	17,32	7	9	11,68	16	18
	T2 ResPC1a4	13,07(13,28)	11	13	11,52(11,67)	18	19	9,4(9,49)	25	28
	T2 PC1a4	13,07(13,28)	5	7	11,52(11,67)	5	8	9,4(9,49)	13	15
	Esperado		3	3,16		6	6,32		15	15,8

Tabela 8 - Comparação de FLC gráfico T^2 matriz covariância usual KDE (linhas 1 a 12) x MVN (linhas 13,14)

Fixado o LSC (22,08 conforme linha 9/coluna2 da Tabela 8) da estatística $\boldsymbol{T^2}$ via KDE (N-300-h/2), foi feita uma análise das observações fora do limite de controle para a taxa de falsos alarmes de 1% (mais rígida) e para a matriz de covariância usual (uma vez que há pouca diferença entre a forma de estimar a covariância quando se lida com os resíduos do modelo ARIMA-Garch das PCs). O gráfico T^2deste cenário para a fase I é exibido na Figura 8 e os valores da contribuição de cada variável (resíduo pc1 a pc4) segundo a equação (5) estão na Tabela 10.

Da análise do peso da cada variável (piezômetro) nas componentes principais Tabela 1 nota-se que os piezômetros que tem maior influencia em pc1 são p_3, p_5, p_6 e p_7 e que estas variáveis são as mais altamente correlacionadas. Uma interpretação para pc1 em termos de sua posição na barragem pode ser que a maior parte da variabilidade deste conjunto de instrumentos é devida a sua localização antes de uma cortina de injeção de concreto e em elevações inferiores, conforme Figura 3, mais propensa a subpressões.

Por exemplo, para a observação 14, indicada como fora de controle, a maior contribuição é de res_pc1, as variáveis que tem maior peso em $pc1$ são p_3, p_5, p_6 e p_7 e ao se analisar a 14ª observação destes instrumentos observa-se uma nítida mudança de comportamento dos dados de todas estas variáveis, sobretudo em p_3. Desta forma, o modelo desenvolvido reduziu a

análise de 7 variáveis a análise de uma variável (estatística T^2), entretanto, isto não impediu a identificação de quando e onde elas ocorreram.

		1%			2%			5%		
Cov. Dif. Sucessiv.		LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II
T2 Resíduos PC1 a PC4 297 obs.	KDE-E-100	19,69	6	7	18,15	9	10	12,98	18	21
	KDE-E-200	24,22	4	5	18,89	8	9	14,05	16	19
	KDE-E-300	25,54	4	5	19,3	6	7	14,57	16	19
	KDE-N-100	20,31	5	6	18,28	9	10	13,22	17	20
	KDE-N-200	24,98	4	5	18,94	7	8	14,41	16	19
	KDE-N-300	29,03	4	5	19,32	6	7	14,8	16	18
	KDE-E-100 h/2	19,36	6	7	18,37	9	10	12,93	19	22
	KDE-E-200 h/2	24,19	5	6	18,83	8	9	13,95	16	19
	KDE-E-300 h/2	28,67	4	5	18,97	6	7	14,77	16	18
	KDE-N-100 h/2	19,46	6	7	18,01	10	11	13,16	17	20
	KDE-N-200 h/2	24,43	4	5	18,88	8	9	14,29	16	19
	KDE-N-300 h/2	29,12	4	5	19,17	6	7	14,99	16	18
	T2 ResPC1a4	13,07	18	20	11,52	22	26	9,4	35	39
	T2 PC1a4	13,07(13,28)	298	315	11,52(11,67)	298	317	9,4(9,49)	298	317
	Esperado		3	3,16		6	6,32		15	15,8

Tabela 9 - Comparação de FLC gráfico T^2 matriz covariância diferenças sucessivas KDE (linhas 1 a 12) x MVN (linhas 13,14)

O uso do KDE aumentou o LSC e diminui a taxa de falsos alarmes e facilitou a análise destes dados fora de controle que, por sua vez, sempre estiveram associados a alguma mudança de comportamento de um e nunca mais do que um instrumento naquela vizinhança com imediata retomada do controle. Isto permite dizer que estatisticamente a variação global do conjunto de instrumentos sempre se manteve sob controle.

Observação	*Valor de* T^2	*RES_PC1*	*RES_PC2*	*RES_PC3*	*RES_PC4*
14	34,0579	14,2925	2,54581	4,86489	2,6689
191	23,7724	8,34781	0,992947	1,35859	15,8347
249	50,9995	3,23206	12,419	4,66065	4,1861
265	22,2737	0,0209662	3,89646	4,82164	0,692861

Tabela 10 - Decomposição T^2 - Contribuição relativa para o valor de T^2 - valores de d_i

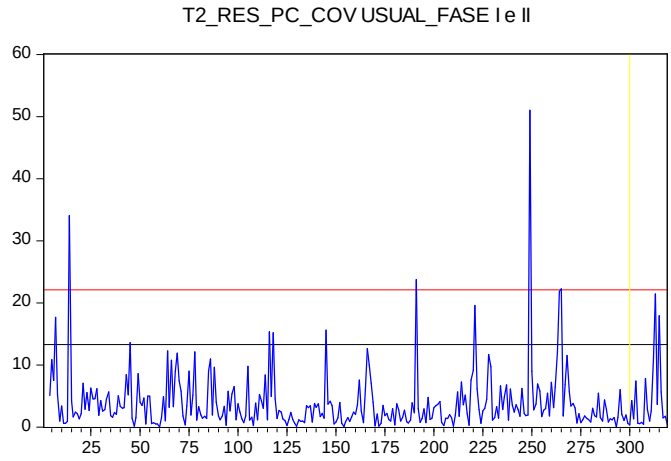

Figura 8 - Gráfico T^2 resíduos das PCs - matriz covariância usual - fases I e II - LSC MVN(preto) e LSC KDE(vermelho)

5. CONCLUSÕES

Neste artigo buscou-se estabelecer valores limites em gráficos de controle multivariados de um conjunto de resíduos de componentes principais de dados de um tipo de instrumento de monitoramento de barragens. As componentes principais foram úteis para a redução da massa de dados e a obtenção de novas variáveis não correlacionadas. A modelagem ARIMA-Garch das componentes serviu para retirar uma parte sistemática de variação (possivelmente originária de causas comuns de variação como a variação térmica) dos dados e os limites de controle foram aplicados aos resíduos deste modelo.

Os resíduos não apresentavam a distribuição normal multivariada então a técnica não paramétrica *kde* foi comparada com os resultados de observações fora do limite de controle do gráfico T^2(que presume MVN) em diversos cenários. Em todos eles a metodologia proposta mostrou desempenho superior em relação ao número médio esperado de falsos alarmes considerando uma base de dados sob controle, tanto na fase I quanto na fase II em relação ao gráfico tradicional MVN.

A combinação de *pca* e *kde* se mostra promissora para estimar quantis específicos de um gráfico de controle multivariado no sentido de reduzir a quantidade de dados, diminuir a taxa de falsos alarmes e não perder a capacidade de detectar mudanças de comportamento em um conjunto de dados específico, produzindo uma análise da qualidade sob uma ótica mais global.

6. REFERÊNCIAS

1. USACE. **Enginnering and Design - Instrumentation of Embankment Dams and Levees**. U.S.Army Corps of Engineers. Washington, DC. 1995.

2. MONTGOMERY, D. C. **Introdução ao Controle Estatístico de Qualidade**. 4ª. ed. Rio de Janeiro-RJ: LTC, 2013.

3. RYAN, T. P. **Statistical Methods for Quality Improvement**. Hoboken-NJ: John Wiley & Sons, 2011.

4. CHENG, L.; ZHENG, D. Two online dam safety monitoring models based on the process of extracting environmental effect. **Advances in Engineering Software**, p. 48-56, 2013.

5. PHALADIGANON, P. et al. Principal component analysis-based control charts for multivariate nonnormal distributions. **Expert Systems with Applications, v.40**, p. 3044-3054, 2013.

6. LIANG, J. Multivariate statistical process monitoring using kernel density estimation. **Dev. Chem. Eng. Mineral Process, v. 13**, p. 185-192, 2005.

7. CHOU, Y.; MASON, R. L.; YOUNG, J. C. The control chart for indivual observations from a multivariate non-normal distribution. **Communications in statistics - Theory and methods**, p. 1937-1949, 2001.

8. VERMAAT, M. B. et al. A comparison of Shewhart individuals control charts based on normal, non-parametric, and extreme-value theory. **Quality and reliability engineering international, v.19**, p. 337-353, 2003.

9. GU, C. et al. Singular value diagnosis in dam safety monitoring effect values. **Science China Technological Sciences, v. 54, nº 5**, May 2011. 1169-1176.

10. YU, H. et al. Multivariate analysis in dam monitoring data with PCA. **Science China Technological Sciences, v. 53, n.4**, April 2010. 1088-1097.

11. SAMOHYL, R. W. **Controle estatístico de qualidade**. Rio de Janeiro: Elsevier, 2009.

12. OAKLAND, J. S. **Statistical Process Control, 6a.ed.** Oxford: Elselvier, 2008.

13. CHAKRABORTI, S. Nonparametric (distribution-free) quality control charts. **Enclyclopedia of Statiscal Sciences - John Wiley & Sons, Inc**, 2011.

14. MASON, R. L.; YOUNG, J. C. **Multivariate statistical process control with industrial aplications**. Philadelphia: American Statistical Association - Society for Industrial and Applied Mathematics, 2002.

15. JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis - 6th. ed.** Upper Saddle River - NJ: Pearson Prentice Hall, 2007.

16. TRACY, N. D.; YOUNG, J. C.; MASON, R. L. Multivariate Control Charts for indiividual observations. **Journal of quality technology**, 1992.

17. HOLMES, D. S.; MERGEN, A. E. Improving the performance of the T2 Control Chart. **Quality Engineering, v.5**, 1993.

18. RUNGER, G. C.; ALT, F. B.; MONTGOMERY, D. C. Contributors to a multivariate statistical process

control signal. **Communications in statistics - Theory and Methods**, v. 25, 1996. ISSN 10.

19. KRUGER, U.; XIE, L. **Statistical monitoring of complex multivariate process with applications in industrial process control**. Chichester: John Wiley & Sons., 2012.

20. NEDUSHAN, B. A. **Multivariate statistical analysis of monitoring data for concrete dams**. Montreal: Tese de Doutorado - Department of Civil Engineering and Apllied Mechanics, 2002.

21. FUNARO, T. C. **Estabelecimento estatístico de valores de controle para instrumentação de barragens de terra - Estudo de caso das barragens de Emborcação e Piau**. Ouro Preto: Dissertação de Mestrado Profissional em Engenharia Geotécnicas da UFPOF, 2007.

22. JACKSON, J. E. **A user's guide to principal components**. New York: John Wiley & Sons, Inc., 1991.

23. JACKSON, J. E. Multivariate quality control. **Communications in statistics - Theory and Methods, v. 14:11**, p. 2657-2688, 1985.

24. MORETTIN, P. A.; TOLOI, C. M. **Análise de séries temporais, 2ª edição rev. e ampl.** São Paulo: Edgard Blucher, 2006.

25. BOX, G. E. P.; JENKINS, G. W.; REINSEL, G. C. **Time Series Analysis Forecasting and Control**. Hoboken-NJ: John Wiley & Sons, Inc., 2008.

26. ENGLE, R. F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom inflation. **Econometrica**, v. 50, p. 987-1007, Julho 1982.

27. BOLLERSLEV, T. Generalized Autorregresive Conditional Heteroscedasticity. **Journal of Econometrics**, v. 31, p. 307-326, 1986.

28. GUJARATI, D. N.; PORTER, D. C. **Econometría**. 5ª Edição. ed. México: McGRAW-HILL/INTERAMERICANA, 2010.

29. HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric statistical methods, 3rd.ed.** New Jersy: John Wiley & Sons, 2014.

30. WAND, M. P.; JONES, M. C. **Kernel smoothing**. New York: Chapmann & Hall, 1995.