

UNIVERSIDADE FEDERAL DO PARANÁ

EMERSON LAZZAROTTO

OBTENÇÃO DE LIMITES ESTATÍSTICOS DE CONTROLE EM GRÁFICOS DE
CONTROLE UNIVARIADOS E MULTIVARIADOS APLICADOS A DADOS DE
INSTRUMENTAÇÃO DE BARRAGENS

CURITIBA

2016

EMERSON LAZZAROTTO

OBTENÇÃO DE LIMITES ESTATÍSTICOS DE CONTROLE EM GRÁFICOS DE
CONTROLE UNIVARIADOS E MULTIVARIADOS APLICADOS A DADOS DE
INSTRUMENTAÇÃO DE BARRAGENS

Tese apresentada como requisito parcial à obtenção do título de Doutor em Métodos Numéricos ao Programa de Pós-Graduação de Métodos Numéricos em Engenharia, na Área de Concentração de Programação Matemática e na Linha de Pesquisa em Métodos Estatísticos Aplicados à Engenharia, da Universidade Federal do Paraná – UFPR.

Orientadora: Prof.^a. Dra. Liliana Madalena Gramani

Coorientador: Prof. Dr. Anselmo Chaves Neto

Coorientador: Prof. Dr. Luiz Albino Teixeira Junior

CURITIBA

2016

Lazzarotto, Emerson

Obtenção de limites estatísticos de controle em gráficos de controle e univariados e multivariados aplicados a dados de instrumentação de barragens / Emerson Lazzarotto. – Curitiba, 2016.

218 f. : il.; tabs.

Tese (doutorado) – Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

Orientadora: Liliana Madalena Gramani

Coorientadores: Anselmo Chaves Neto, Luiz albino Teixeira Junior

Bibliografia: p.199-209

1. Controle estatístico de qualidade. 2. Monitoramento ambiental.
3. Modelagem de processos. I. Gramani, Liliana Madalena. II. Chaves Neto, Anselmo. III. Teixeira Junior, Luiz Albino. IV. Título.

CDD 526.982



Ministério da Educação
Universidade Federal do Paraná
Setor de Tecnologia / Setor de Ciências Exatas
Departamento de Construção Civil / Departamento de Matemática/ Departamento
de Engenharia de Produção.
Programa de Pós-Graduação em Métodos Numéricos em Engenharia –
PPGMNE/UFPR.



TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de EMERSON LAZZAROTTO, intitulada: "OBTENÇÃO DE LIMITES ESTATÍSTICOS DE CONTROLE EM GRÁFICOS DE CONTROLE UNIVARIADOS E MULTIVARIADOS APLICADOS A DADOS DE INSTRUMENTAÇÃO DE BARRAGENS", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVADO.

Curitiba, 11 de Março de 2016.

Prof LILIANA MADALENA GRAMANI (UFPR)
(Presidente da Banca Examinadora)

Prof INÁCIO ANDRUSKI GUIMARÃES (UTFPR)

Prof ISABELLA ANDRECZEWSKI CHAVES (UFPR)

Prof JAIR MENDES MARQUES (UFPR)

Prof SACHIKO ARAKI LIRA (UFPR)

Prof SELMO KUPERMAN (USP)

DEDICATÓRIA

À minha família.

A memória de meu pai e meu irmão.

Aos amigos e amigas pelo companheirismo e confiança.

Aos professores de toda minha carreira estudantil.

AGRADECIMENTOS

A Deus por me dar a vida e a capacidade física e intelectual de chegar até aqui.

À minha família, hoje formada pela minha mãe (Rosália) e minha irmã (Nirdes), base sólida, sem a qual também não poderia chegar até aqui.

À minha orientadora, Prof.^a. Dra. Liliana Madalena Gramani, pela orientação prestada, confiança depositada, liberdade e tratamento dispensado. Além disso, pelo seu permanente trabalho e confiança na turma DINTER fora de sede.

Ao meu coorientador prof. Dr. Anselmo Chaves Neto, pelo conhecimento passado, profissionalismo demonstrado, pela colaboração, disponibilidade e valiosas sugestões, além da amizade. Ambos serão lembrados pelo comportamento ético e visão do mundo acadêmico que me transmitiram.

Ao CEASB-FPTI/BR e à Itaipu Binacional por viabilizarem economicamente a abertura de uma turma DINTER fora de sede do PPGMNE e pela cessão de dados e autorização para divulgação. Em especial ao engenheiro Dimilson Pinto Coelho – Diretor do CEASB, por sua disponibilidade, humildade e suas “aulas” de Itaipu.

Ao colega de turma Engenheiro (Dr.) Geraldo Carvalho Brito Junior, idealizador desta turma fora de sede, pelo seu trabalho e amizade. A todos os demais colegas de turma, em especial, Fernando, Orlando e Cleverson.

Ao coorientador professor Dr. Luiz Albino Teixeira Junior, pela sua colaboração em todos os sentidos, sobretudo, com ideias, discussões esclarecedoras e confiança transmitida.

Ao colegiado do curso de matemática da UNIOESTE – campus de Foz de Iguaçu - pela autorização para afastamento em uma parte do desenvolvimento deste trabalho, fator fundamental para sua conclusão, em especial aos colegas Claiton Petris Massarolo e Luciano Panek, e aos servidores da UNIOESTE.

Ao povo do Estado do Paraná que ‘financiou’ meu afastamento e é quem, em última instância, permite a realização deste trabalho.

À minha companheira, Ellen Vieira, por ser um importante ponto de apoio durante o desenvolvimento deste trabalho.

À CAPES e à Fundação Araucária pelo auxílio financeiro durante uma parte deste doutorado.

EPÍGRAFE

Alcançar a conquista é mais importante do que aquilo que é conquistado.
Sobretudo, quando alicerçada sobre valores éticos e pacíficos.

Emerson Lazzarotto

RESUMO

Barragens instrumentadas geram uma enorme base de dados provenientes do monitoramento. Porém, dados não são informações e, então, precisam ser trabalhados para se obter informações. A correta interpretação e avaliação da informação contida nos dados podem ser extremamente úteis para o acompanhamento do período operacional de uma barragem. O controle estatístico de qualidade univariado e multivariado é uma abordagem que pode contribuir com a descoberta de conhecimento em dados de instrumentos de monitoramento de barragens, sobretudo quando há correlação e quando não se tem informação sobre o comportamento das séries temporais de monitoramento dos instrumentos. Enquanto que gráficos de controle univariados são aplicados, há quase um século, no controle de características de qualidade, propostas multivariadas e modelos aplicados em segurança de barragens têm ainda poucos trabalhos publicados e têm se mostrado promissoras. Por outro lado, no caso univariado, o avanço da informática permitiu a execução de novas técnicas para modelar e realizar previsões em séries temporais, em particular, a combinação de técnicas com modelos ARIMA, ARIMAX-GARCH, decomposição *Wavelet* e Redes Neurais Artificiais tem mostrado melhoria no desempenho da previsão. O objetivo principal deste estudo é aplicar técnicas estatísticas univariadas e multivariadas, combinadas com outras técnicas paramétricas e não paramétricas aos dados de instrumentação de barragens a fim de desenvolver modelos capazes de estabelecer valores estatísticos de controle e previsões às séries temporais de leituras de instrumentos para a avaliação do desempenho futuro de uma barragem. Os modelos desenvolvidos foram testados e validados nos dados de instrumentos do trecho E da Usina Hidrelétrica de Itaipu Binacional, localizada no Rio Paraná, entre o Brasil e o Paraguai. Os resultados mostraram que o gráfico multivariado T^2 das componentes principais combinado com os modelos ARIMAX-GARCH e técnicas não paramétricas permitem manter as taxas de falsos alarmes muito mais próximas do valor esperado em relação aos gráficos univariados e T^2 multivariados, quando se presume normalidade nos dados. Além disso, a utilização de modelos híbridos, que combinam a decomposição *Wavelet*, a modelagem ARIMA-GARCH e Redes Neurais Artificiais pode melhorar o desempenho da previsão da série temporal da leitura de um instrumento comparativamente à utilização individualizada destas técnicas.

Palavras-chave: Controle Estatístico de Qualidade; monitoramento numérico de barragens; gráficos de controle; Séries Temporais; modelos ARIMA-GARCH; Redes Neurais; Decomposição *Wavelet*; técnicas não paramétricas.

ABSTRACT

Dams with instruments generate a huge base of data from the monitoring. However, data is not information, and then need to be work to obtain information. The correct interpretation and evaluation of the information contained in the data can be extremely useful for monitoring the operational period of a dam. The univariate and multivariate statistical quality control is an approach that can contribute to knowledge discovery in data dam monitoring instruments, especially when there is correlation and when there is no information on the behavior of time series of monitoring instruments. While univariate control charts are applied for nearly a century in the quality control of characteristics, multivariate proposals and models applied in dam safety still have few papers published and have proven to be promising. On the other hand, in the univariate case, the advance of computer technology has allowed the implementation of new techniques to model and make predictions in time series, in particular, the combination of techniques ARIMA models ARIMAX-GARCH, Wavelet decomposition and artificial neural networks has shown improved performance of the forecast. The main goal of this study is to apply univariate and multivariate statistics techniques, combined with other parametric and nonparametric techniques to dam instrumentation data to develop models able to establish statistical control values and forecast the time series of instrument readings for assessment of the future performance of a dam. The developed models have been tested and validated on data from instruments of the E section of the Binational Itaipu hydroelectric plant, located on the Paraná River, between Brazil and Paraguay. The results showed that the multivariate graphic T^2 of the principal components combined with ARIMAX-GARCH and nonparametric techniques allows to maintain the rates of false alarms much closer of the expected value in relation to the univariate graphs and T^2 multivariate, when data are assumed to be normal. Furthermore, the use of hybrid models that combine the Wavelet decomposition, ARIMA-GARCH modeling and artificial neural networks can improve time-series forecasting performance of an instrument reading compared to the individual use of these techniques.

Key-words: statistical quality control; numerical monitoring instruments; control charts; time series; ARIMA-GARCH models; neural networks; Wavelet decomposition; nonparametric techniques.

LISTA DE FIGURAS

FIGURA 2.1 - Acesso à energia e índice de desenvolvimento humano.....	25
FIGURA 2.2 - Esquema básico de uma usina hidrelétrica	26
FIGURA 2.3 – Curva de probabilidade de falhas de uma barragem com o tempo....	28
FIGURA 2.4 - Produção energética do Brasil em 2013 por fontes	28
FIGURA 2.5 - Esquema de funcionamento de pêndulos diretos e invertidos em Itaipu	40
FIGURA 2.6 - Esquema de um extensômetro múltiplo em Itaipu	41
FIGURA 2.7 – Distribuições das dimensões críticas para as transmissões	42
FIGURA 2.8 – Ciclo de Deming.....	43
FIGURA 2.9 - Um típico gráfico de controle	44
FIGURA 2.10 - Esquema de um gráfico de controle	45
FIGURA 2.11 - Diferença entre distância estatística e distância Euclidiana	52
FIGURA 2.12 - Esquema gráfico de duas variáveis altamente correlacionadas	54
FIGURA 2.13 - Elipse de controle bivariada e região de controle retangular de 2 gráficos de Shewhart.....	56
FIGURA 2.14 - Variável cíclica no tempo	65
FIGURA 2.15 - Interpretação geométrica de componentes principais para $p=2$	70
FIGURA 2.16 - Um típico <i>scree plot</i>	72
FIGURA 2.17 - Componentes principais e uma elipse com distância constante.....	73
FIGURA 2.18 - Diagrama de um modelo de componentes principais e do seu subespaço complementar residual.....	74
FIGURA 2.19 - Estágios do ciclo iterativo da construção de um modelo de séries temporais.....	80
FIGURA 2.20 – Uma Wavelet tipo chapéu mexicano.....	91
FIGURA 2.21 – Uma Wavelet de Morlet	91
FIGURA 2.22 - Decomposição Wavelet de nível 2 da série temporal de um instrumento de Itaipu	92
FIGURA 2.23 - Arquitetura de uma RNA <i>feedforward</i> com uma camada oculta e uma saída	98
FIGURA 2.24 - Esquema da reamostragem <i>Bootstrap</i>	105
FIGURA 2.25 - Comparação de diferentes valores de h para uma amostra da distribuição normal padrão	110
FIGURA 2.26 - Efeito de diferentes valores de h no KDE	111
FIGURA 2.27 - Esquema de procedimento de controle estatístico multivariado	120
FIGURA 2.28 - Evolução de técnicas de controle estatístico bivariado.....	127
FIGURA 2.29 - Painel de alerta para o bloco Chave F13/14 em Itaipu	133
FIGURA 3.1 - Arranjo geral do projeto de Itaipu.....	139
FIGURA 3.2 - Barragem da Usina de Itaipu e o Trecho E em destaque	140

FIGURA 3.3 - Parte da instrumentação do bloco E-6.....	140
FIGURA 3.4 - Instrumentação do Bloco E-6 em Itaipu.....	142
FIGURA 3.5 - Resposta defasada de um piezômetro em relação ao nível do reservatório	143
FIGURA 3.6 - Local de leituras dos piezômetros 28 e 29 no trecho E	144
FIGURA 3.7 - Fluxograma para a abordagem 7 do QUADRO 3.2.....	149
FIGURA 3.8 - Seção do bloco 6 do trecho E de Itaipu com instrumentos p1 até p7 em destaque.....	154
FIGURA 3.9 - Fluxograma do método empregado na seção 3.3.2.2	156
FIGURA 3.10 - Fluxograma das etapas da metodologia para a subseção 3.3.2.3..	161
FIGURA 4.1 - Gráficos da série original e suas componentes <i>Wavelet</i>	163
FIGURA 4.2 - Desempenho da melhor abordagem no período de validação	165
FIGURA 4.3 - Série temporal da vazão média mensal em Itaipu.....	167
FIGURA 4.4 - Evolução mensal do erro absoluto percentual (APE) na amostra de teste do modelo CL2 (linha azul) e do modelo proposto (linha vermelha).....	168
FIGURA 4.5 – Gráfico T^2 para quatro componentes principais retidas incluindo dados das fases I e II	173
FIGURA 4.6 - Elipse de controle dos escores das duas primeiras componentes de quatro componentes principais retidas.....	173
FIGURA 4.7- Gráfico da estatística Q considerando k=4 componentes principais nas fases I e II.....	173
FIGURA 4.8 - Correlograma de pc1 (FAC e FACP).....	177
FIGURA 4.9 - Correlograma dos resíduos e resíduos quadráticos de pc1	178
FIGURA 4.10 - Exemplo de kernel funções de estimativa da densidade de probabilidade dos valores de T^2	180
FIGURA 4.11 – Gráfico do escore T^2 dos resíduos das PC's - matriz covariância usual - fases I e II - LSC MVN(preto) e LSC KDE(vermelho) em relação a observação (amostra)	183
FIGURA 4.12 - Instrumentos mais importantes da CP z_1	187
FIGURA 4.13 - Instrumentos mais importantes da CP z_2 (acima) e da CP z_3 (abaixo)	188
FIGURA 4.14 - Instrumentos mais importantes da CP z_4 (acima) e da CP z_5 (abaixo)	189
FIGURA 4.15 - Gráfico T^2 da série $t2_{res}$, LSC MVN x KDE	194

LISTA DE TABELAS

TABELA 2.1 – Matriz da capacidade instalada de geração de energia elétrica no Brasil- Janeiro 2016 (MW).....	27
TABELA 2.2 - Total de instrumentos de auscultação das estruturas civis de Itaipu..	35
TABELA 3.1 - Separação da amostra	157
TABELA 4.1 - RNA' s ótimas encontradas para as abordagens estabelecidas	163
TABELA 4.2 - Modelos <i>ARIMA-GARCH</i> ótimos encontrados para as componentes	164
TABELA 4.3 - Resultados finais do desempenho das previsões por abordagem ...	164
TABELA 4.4 - Valores do MAPE nas amostras de treinamento, validação e teste .	167
TABELA 4.5 - Valores do MAE nas amostras de treinamento, validação e teste....	167
TABELA 4.6 - Observações fora dos limites de controle (FLC) para gráficos univariados	169
TABELA 4.7 - Número de observações FLC para cada matriz de covariância para o gráfico <i>T2</i>	170
TABELA 4.8 - Variabilidade Explicada pelas componentes principais	172
TABELA 4.9 - Autovetores da matriz de covariância amostral <i>S</i> (equação (2.1.33))	172
TABELA 4.10 - Resultados do Gráfico <i>T2</i> das <i>k</i> CP's obtido a partir da retenção de <i>k</i> componentes principais.....	172
TABELA 4.11 - Resultados do Gráfico Q das <i>p-k</i> componentes principais	172
TABELA 4.12 - Análise das componentes principais de <i>p1</i> a <i>p7</i> autovalores (esquerda) e autovetores (direita)	176
TABELA 4.13 - Análise das componentes principais de <i>p1</i> a <i>p7</i> - Correlações	176
TABELA 4.14 - Análise das componentes principais de <i>p1</i> a <i>p7</i> - Teste de normalidade.....	176
TABELA 4.15 - Teste de normalidade Jarque-Bera aplicado aos dados originais padronizados (<i>p1</i> a <i>p7</i>).....	176
TABELA 4.16 - Resultados da modelagem <i>ARIMA-Garch</i> de <i>pc1</i>	178
TABELA 4.17 - Teste de normalidade Jarque-Bera para as componentes retidas (fases I e II)	179
TABELA 4.18 - Teste de normalidade Jarque-Bera para os resíduos do modelo <i>ARIMA-GARCH</i> das componentes retidas (fases I e II)	179
TABELA 4.19 - Comparação de FLC gráfico <i>T2</i> matriz de covariância usual KDE (linhas 1-12) x MVN (linhas 13,14)	181
TABELA 4.20 - Comparação de FLC gráfico <i>T2</i> matriz de covariância por diferenças sucessivas KDE (linhas 1- 12) x MVN (linhas 13,14)	182
TABELA 4.21 - Decomposição <i>T2</i> - Contribuição relativa para o valor de <i>T2</i> - valores de <i>di</i>	183
TABELA 4.22 - Resumo da modelagem <i>ARIMA-GARCH</i> da CP <i>z1</i>	186

TABELA 4.23 - Resultados da modelagem para CP z1	190
TABELA 4.24 - Resultados dos melhores modelos encontrados	190
TABELA 4.25 - Teste de Jarque-Bera para normalidade dos resíduos das CP's e das CP's.....	191
TABELA 4.26 - LSC e quantidade de FLC para as fases I e II, $\alpha=1;2$ e 5% para os resíduos e para as CP's	191
TABELA 4.27 - Resultados dos valores do LSC e FLC obtidos via KDE para a série temporal t2_res	192
TABELA 4.28 - Resultados dos valores do LSC e FLC obtidos via KDE para a série temporal t2_pcs	193

LISTA DE QUADROS

QUADRO 2.1 – Alguns termos básicos em barragens.....	24
QUADRO 2.2 - Instrumentos nas estruturas de concreto de Itaipu.....	39
QUADRO 2.3 – Instrumentos nas fundações da barragem de Itaipu.....	40
QUADRO 2.4 – Valores e distribuição da estatística T_2 conforme os parâmetros e a fase	60
QUADRO 2.5 - Valores de Atenção para os piezômetros de barragem de gravidade convencional	131
QUADRO 2.6 - Tempo para análise de causas baseado no nível de atenção atingido	132
QUADRO 2.7 - Condições adotadas no MEF para definir valores de controle no bloco E-6	133
QUADRO 3.1 - Instrumentos localizados no bloco/trecho E6	141
QUADRO 3.2 - Abordagens testadas.....	148
QUADRO 3.3 – Séries temporais utilizadas no trabalho da seção 3.2.2.3.....	158

LISTA DE SIGLAS E ABREVIATURAS

ACP	- Análise de componentes principais
AIC	- <i>Akaike Information Criterion</i>
ANEEL	- Agência Nacional de Energia Elétrica
ARIMA	- Autorregressivo integrado de média móvel
ARIMAX	- ARIMA com variáveis exógenas (ou explicativas)
ARL	- <i>Average Run Length</i>
BDS	- Teste Brock, Dechert, Sheinkman de independência
CEASB	- Centro de Estudos Avançados em Segurança de Barragem
CMS	- Comprimento médio da sequência
CP ou PC	- Componente principal
CUSUM	- Gráfico de controle de somas cumulativas
EWMA (MMEP)	- Gráfico de controle de médias móveis exponencialmente ponderadas
f.d.a.	- Função distribuição acumulada
f.d.p.	- Função densidade de probabilidade
FAC	- Função de autocorrelação
FACP	- Função de autocorrelação parcial
FLC	- Observação fora dos limites de controle
FPTI/BR	- Fundação Parque Tecnológico de Itaipu – Brasil
GARCH	- <i>Generalized Autoregressive Conditional Heteroscedasticity</i>
H_0, H_a	- Hipótese nula e hipótese alternativa
i.i.d.	- Independentes e identicamente distribuídos
ICOLD	- <i>International Committee of Large Dams</i>
JB	- Teste de Jarque-Bera
KDE	- <i>Kernel Density Estimation</i>
LB	- Teste de Ljung-Box
$\mathcal{L}^2(\mathbb{R})$	- Espaço das funções quadrado integráveis
l^2	- Conjunto das sequências infinitas quadrado somáveis
LIC	- Limite inferior de controle
LIE	- Limite inferior de especificação
LSC	- Limite superior de controle
LSE	- Limite superior de especificação
MAE	- <i>Mean absolute error</i> (erro absoluto médio)
MAPE	- <i>Mean absolute percentual error</i> (Erro percentual absoluto médio)
MSE	- <i>Mean square error</i> (Erro quadrático médio)
msnm	- Metros sobre nível do mar
MVN	- Normalidade multivariada
ONS	- Operador Nacional do Sistema Elétrico
PE	- Processo estocástico
PNSB	- Política Nacional de Segurança de Barragem

RMSE	- <i>Root mean square error</i> (raiz do erro quadrático médio)
RNA	- Rede Neural artificial
SPE	- <i>Square prediction error</i> (erro quadrático de previsão)
ST	- Série temporal
UAR	- Unidade de aquisição remota
USACE	- <i>United States American Corporate Engineers</i>

LISTA DE SÍMBOLOS

Σ	- Matriz de covariância populacional
S ou S_1	- Matriz de covariância amostral
σ	- Desvio padrão populacional
s	- Desvio padrão amostral
ρ	- Coeficiente de correlação populacional
X	- Vetor observação (multivariado)
X'	- Vetor observação transposto
\bar{X}	- Vetor de médias amostral (multivariado)
\bar{x}	- Média amostral (univariada)
S_2	- Matriz de covariância por diferenças sucessivas
N	- Distribuição normal
β	- Distribuição beta
χ^2	- Distribuição Qui-quadrado
μ	- Média populacional (univariada)
$\boldsymbol{\mu}$	- Vetor de médias populacional (multivariada)
T^2	- Estatística T^2 de Hotelling
λ_i	- i -ésimo autovalor
F	- Distribuição F de Snedecor
\sim	- Segue uma distribuição...
$\#$	- Cardinalidade de um conjunto
ep	- Erro padrão
t	- Distribuição t
R^2	- Grau de ajustamento de um modelo
x^*	- (Re)Amostra <i>Bootstrap</i>

SUMÁRIO

1	INTRODUÇÃO.....	18
1.1	O PROBLEMA.....	18
1.2	OBJETIVOS.....	19
1.2.1	Objetivo Geral	19
1.2.2	Objetivos Específicos.....	19
1.3	JUSTIFICATIVA	20
1.4	ESTRUTURAÇÃO DA TESE.....	22
2	REVISÃO BIBLIOGRÁFICA.....	23
2.1	FUNDAMENTAÇÃO TEÓRICA.....	23
2.1.1	A Importância das Barragens.....	23
2.1.2	Segurança e Instrumentação de barragens	29
2.1.3	Controle Estatístico de Qualidade.....	41
2.1.3.1	Introdução.....	41
2.1.3.2	Gráficos de controle multivariados.....	50
2.1.3.3	Gráfico de controle T^2	56
2.1.3.4	Gráfico de controle para dados autocorrelacionados.....	62
2.1.4	Análise de Componentes Principais (ACP).....	65
2.1.5	Séries Temporais e Modelagem ARIMAX-GARCH.....	75
2.1.6	Decomposição <i>Wavelet</i> de uma Série Temporal	89
2.1.7	Redes Neurais Artificiais	96
2.1.8	Estatística Não Paramétrica.....	100
2.1.9	<i>Bootstrap</i>	104
2.1.10	Núcleos Estimadores da Densidade (<i>Kernel Density Estimation</i>)	107
2.1.11	Problema de Programação Matemática	112
2.2	REVISÃO DE LITERATURA	113
2.2.1	Gráficos Multivariados, Experiências Internacionais em Barragens	113
2.2.2	Algumas Experiências com ACP.....	117
2.2.3	Gráficos de Controle Não Paramétricos.....	122
2.2.4	Modelos Híbridos	127
2.2.5	Algumas Experiências em Itaipu e Outras Barragens no Brasil.....	129
3	MATERIAL E MÉTODOS.....	137

3.1 MATERIAL	138
3.2 MÉTODOS	145
3.2.1 Caso Univariado.....	145
3.2.1.1 O caso do pêndulo direto.....	145
3.2.1.2 O caso da série da vazão de afluentes no Rio Paraná.....	149
3.2.2 Caso multivariado.....	151
3.2.2.1 Estatísticas $T2$ e Q no estudo dos piezômetros.....	151
3.2.2.2 O caso dos piezômetros do trecho E-6 via KDE.....	152
3.2.2.3 Análise de todo conjunto de instrumentos.....	156
4 RESULTADOS	162
4.1 CASO UNIVARIADO	162
4.1.1 O Caso do Pêndulo Direto.....	162
4.1.2 O Caso da Série da Vazão de Afluentes no Rio Paraná.....	166
4.2 CASO MULTIVARIADO	168
4.2.1 Estatísticas $T2$ e Q no Estudo dos Piezômetros.....	168
4.2.2 O Caso dos Piezômetros do Trecho E6 via KDE.....	175
4.2.3 Análise de Todo Conjunto de Instrumentos.....	185
5 CONCLUSÕES	197
6 TRABALHOS FUTUROS	198
REFERÊNCIAS	199
APÊNDICE A (CNMAC)	210
APÊNDICE B (SMNC)	211
APÊNDICE C (BOARD)	212
APÊNDICE D (SNGB)	213
APÊNDICE E (CMN)	215
APÊNDICE F (IJMP)	216
APÊNDICE G (CADERNOS DO IME)	217
APÊNDICE H (EAAS - IJEAS)	218

1 INTRODUÇÃO

1.1 O PROBLEMA

A construção de uma usina hidrelétrica traz uma série de benefícios para a população, entretanto, tendo em vista a grandiosidade da obra, mesmo se utilizando de diversos procedimentos para assegurar a segurança da barragem é fundamental continuar se investindo em novos conhecimentos e equipamentos capazes de garantir a segurança da população e o bom funcionamento da usina.

A Usina Hidrelétrica Binacional de Itaipu, localizada no Rio Paraná, na divisa entre o Brasil e o Paraguai, em seu projeto de construção, levou em consideração a instalação de mais de dois mil instrumentos para a análise da auscultação da barragem com vistas à garantia da contínua manutenção da operação da usina que causa enorme impacto social e econômico.

Além dos conhecimentos, inerentes à área de engenharia e da capacidade da equipe de engenheiros no acompanhamento e análise da grande massa de dados recebida dos instrumentos diariamente, o uso de técnicas estatísticas pode contribuir na transformação dos dados em informações capazes de aumentar o conhecimento sobre a segurança da barragem, os quais, somados aos anteriores, podem qualificar as tomadas de decisões dos gestores.

Uma questão que, mesmo o mais experiente engenheiro, pode não ter resposta é: de que forma os instrumentos de monitoramento estão relacionados entre si? Quais são os valores de controle do monitoramento da leitura de um dado instrumento? Leituras anômalas de um determinado instrumento indicam mudança no comportamento da barragem, falha no instrumento ou alguma causa temporária de variação? Quais as vantagens e desvantagens de estudar o comportamento de vários instrumentos simultaneamente? Qual abordagem permite a previsão das séries de leitura de um dado instrumento com maior exatidão?

Desta forma, uma série de tarefas em relação à instrumentação precisa ser executada, dentre essas: o estabelecimento de valores de controle baseados no período operacional, identificação dos instrumentos mais importantes na análise da

variabilidade dos dados de leitura, identificação e eventual remoção de valores discrepantes (*outliers*), interpretação dos dados, ajuste de modelos e estabelecimento de alertas de detecção de mudanças estatisticamente significativas nas respostas da barragem. Em resumo, as duas principais demandas que se vislumbram com as séries de dados de leituras de instrumentos de monitoramento são:

1. Determinar limites estatísticos de controle para aquele conjunto de dados e
2. Prever o comportamento futuro de um ou mais instrumentos da forma mais precisa possível.

Do ponto de vista acadêmico, a aplicação em questão pode suscitar as seguintes questões: de que forma se pode aumentar o desempenho (redução do erro) de uma previsão de séries temporais? A combinação de técnicas é viável? O tratamento multivariado pode reduzir o trabalho sem perda de informação relevante e produzir novas descobertas nos dados?

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O principal objetivo deste estudo é aplicar técnicas de análise estatística univariada e multivariada, decomposição *Wavelet*, modelos de Redes Neurais e outras técnicas não paramétricas às séries temporais dos dados de leituras de instrumentos de monitoramento da barragem de uma usina hidrelétrica, para determinar limites estatísticos de controle aos dados e melhorar o desempenho da previsão das séries, no intuito de colaborar com a equipe de segurança de barragem ao tornar suas decisões menos empíricas.

1.2.2 Objetivos Específicos

a) Avaliar a viabilidade dos métodos existentes de gráficos de controle aplicados em monitoramento de instrumentação de barragens;

- b) Propor um novo método para construção de gráficos de controle univariados e multivariados aplicados ao controle estatístico de qualidade de monitoramento de instrumentação de barragem;
- c) Avaliar o desempenho do método proposto e comparar seu desempenho com outras técnicas clássicas no intuito de 'melhorar' a qualidade das previsões e dos limites de controle;
- d) Identificar e interpretar fatores explicativos da maior parte da variabilidade dos dados provenientes da leitura e eventuais instrumentos que não apresentam confiabilidade na leitura;
- e) Estabelecer, com base em procedimentos estatísticos univariados e multivariados (paramétricos e/ou não paramétricos), valores limites (cartas de controle) para séries temporais de leituras de instrumentos capazes de auxiliar no fornecimento de alertas precoces de mudanças no comportamento da barragem.

1.3 JUSTIFICATIVA

As barragens são barreiras ou estruturas que cruzam córregos, rios ou canais para confinar e controlar o fluxo de água e são geralmente usadas para um ou mais fins específicos, como fornecimento de água, geração de energia hidrelétrica e irrigação. Dentre outros requisitos administrativos, as barragens devem ter instrumentação adequada e monitoramento de desempenho. A manutenção de barragens seguras é garantia do desenvolvimento e a gestão sustentável dos recursos hídricos (ICOLD - CIGB, 2008).

O controle do comportamento estrutural de uma grande obra de engenharia, como uma barragem de usina hidrelétrica é, normalmente, feito por meio de inspeções visuais e de instrumentação. As inspeções visuais são mais utilizadas na detecção de acidentes e incidentes. Entretanto isto não diminui a importância da instrumentação que, ou é insuficiente, ou não é devidamente explorada. Em especial, o melhor uso da instrumentação pode minimizar acidentes e incidentes, estabelecer normas de operação, propor medidas de conservação e permitir o aprofundamento dos conhecimentos existentes proporcionando a melhoria da qualidade e do tempo de vida da obra (SILVEIRA, 2003).

Em aplicações de monitoramento e controle de falhas, o ajuste dos níveis de alarme é muito importante (MILETIC, *et al.*, 2004, p. 834). A literatura dos congressos internacionais de grandes barragens enfatiza a necessidade da manutenção de uma cultura de segurança e de sistemas efetivos de gerenciamento de segurança.

Em Itaipu Binacional (2002) são propostas estimativas estatísticas de deslocamentos na usina através da análise de regressão das leituras efetuadas durante a vida útil da barragem, a qual foi concebida para funcionar em longo prazo (ao menos 100 anos) e o (re)estabelecimento de 'níveis de atenção' após o término da fase de enchimento do reservatório em 1984 e nos anos de 2014, 2044 e 2084.

A responsabilidade pela resolução de problemas e a melhoria de um processo não deve recair apenas sob um grupo ou departamento, mas deve ser compartilhada por todos os envolvidos. Esse é um dos princípios básicos da qualidade total. Os problemas na operação de um processo raramente são devidos a uma única causa, mas, sim, a uma combinação de fatores. Deve ser dada ênfase na prevenção e na manutenção e na melhoria da qualidade de um produto ou processo de uma empresa (OAKLAND, 2008).

O gráfico de controle é uma das formas de controle estatístico de qualidade mais utilizada e pode ser muito útil no controle da instrumentação. A facilidade do fornecimento *on-line* de dados com elevada frequência disponibiliza uma enorme massa de dados que gera uma série de gráficos de controle, entretanto, é necessário interpretar estes dados para que produzam conhecimento adicional, além de meras séries históricas de dados.

Há relativamente poucos trabalhos publicados relacionados à identificação e a interpretação de fatores de variabilidade em instrumentação e monitoramento de barragens de usina hidrelétricas tratando as bases de dados de instrumentação. Os trabalhos existentes não respondem objetivamente a perguntas, tais como: qual a melhor metodologia para lidar com cartas de controle de leitura de instrumentos de auscultação de barragens (univariada ou multivariada, paramétrico x não paramétrico)? Como se pode melhorar o desempenho de um gráfico de controle neste contexto? O uso de técnicas híbridas que levem em conta o comportamento linear e não linear pode melhorar o desempenho da previsão de uma série?

1.4 ESTRUTURAÇÃO DA TESE

Esta tese está estruturada em 6 capítulos e 8 Apêndices, descritos a seguir. O capítulo 1, em questão, apresenta o problema abordado, os objetivos a serem alcançados e as justificativas da escolha do tema.

O capítulo 2 apresenta uma revisão bibliográfica do tema proposto neste trabalho, ou seja, alguns métodos de modelagem de séries temporais e gráficos de controle univariados e multivariados aplicados a dados de monitoramento de instrumentos de barragens de usinas hidrelétricas. Para tanto, na seção 2.1 foi feita uma pesquisa sobre a importância e a instrumentação de barragens, controle estatístico de qualidade, modelagem ARIMA de séries temporais, Decomposição *Wavelet*, Redes Neurais e gráficos de controle multivariados e das técnicas não paramétricas *Bootstrap* e KDE e na seção 2.2 tem-se uma revisão da literatura científica acerca do tema.

O Capítulo 3 descreve os dados utilizados e uma breve justificativa da escolha dos métodos e a sistemática empregada em cada etapa do trabalho desenvolvido. O capítulo 4 apresenta os resultados obtidos para cada trabalho univariado ou multivariado. O capítulo 5 contempla as considerações finais e o capítulo 6, algumas sugestões de trabalhos futuros. Em seguida são listadas as referências bibliográficas utilizadas e os apêndices contêm um resumo dos trabalhos já realizados até o momento.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo, as subseções 2.1.1, 2.1.2 e 2.1.3 realizam um detalhamento do contexto onde o problema abordado neste trabalho está inserido. Nas subseções 2.1.4 a 2.1.11 os métodos selecionados são apresentados mais detalhadamente e a seção 2.2 contém uma revisão de artigos científicos relacionados ao problema e aos métodos utilizados.

2.1 FUNDAMENTAÇÃO TEÓRICA

2.1.1 A Importância das Barragens

A água é elemento essencial à vida, portanto, em todos os seus aspectos é básica para a civilização humana. A necessidade de dispor permanentemente de água em muitas situações se confronta com a escassez provocada pelas inconstâncias do ciclo hidrológico. Para superar esta dificuldade foram criadas formas para o armazenamento e a distribuição controlada da água ao longo do tempo, utilizando obstáculos artificiais aos cursos de água: as barragens (ICOLD - CIGB, 2008).

As barragens e reservatórios têm oferecido ao longo do tempo uma série de benefícios e contribuições para o desenvolvimento sustentável dos recursos hídricos e energéticos e a Engenharia Civil para o progresso e a melhoria da qualidade de vida das pessoas (CIGB-ICOLD, 2008).

Na Organização das Nações Unidas, 189 países colocaram como metas de desenvolvimento do milênio a serem alcançadas até 2015, uma série de objetivos de melhoria social e econômica, dentre estes: diminuir a pobreza nos países mais pobres, globalizar valores e direitos humanos universalmente aceitos, erradicar extrema pobreza e fome, alcançar uma educação primária universal, promover a igualdade de gêneros, reduzir a mortalidade infantil, melhorar a saúde materna, combater a AIDS/HIV, malária e outras doenças, assegurar a estabilidade do meio ambiente e desenvolver parcerias globais para o desenvolvimento. Embora o acesso à energia não seja uma meta propriamente dita, é evidente que o acesso e o fornecimento adequado de energia são elementos cruciais para alcançar estes objetivos (XXIII ICOLD QUESTION 88, 2009).

No QUADRO 2.1 são apresentadas as definições de alguns termos fundamentais em barragens, segundo (ICOLD - CIGB, 2008).

QUADRO 2.1 – Alguns termos básicos em barragens

TERMO	EXPLICAÇÃO
Montante	Ponto referencial visualizado pelo observador que olha em direção à nascente de um curso de água (águas acima), ou seja, a nascente é o ponto mais a montante do rio.
Jusante	Ponto referencial visualizado pelo observador que olha em direção à foz de um curso de água (águas abaixo), ou seja, a foz é o ponto mais a jusante de um rio.
Margem direita do reservatório (ou rio ou qualquer curso de água)	Situada à direita do observador que olha para jusante (para onde corre o fluxo de água, águas abaixo).
Margem esquerda do reservatório	Situada à esquerda do observador que olha para jusante.

FONTE: ICOLD-CIGB (2008)

A água é, e continuará sendo, recurso vital para a civilização humana. O suprimento adequado e seguro de água são componente essencial da nossa saúde, nosso ambiente, comunidades e economia. A mudança climática cada vez mais frequente e o aumento da população mundial que aumenta a demanda por água para fins domésticos, agrícolas e industriais exigem cada vez mais a preocupação com o armazenamento de água. Neste contexto as barragens, desde a antiguidade, exercem e continuarão a exercer papel crucial na história da nossa civilização.

A FIGURA 2.1 apresenta uma relação entre a questão do acesso à energia e o Índice de Desenvolvimento Humano no Brasil no ano 2000. Isto confirma que a energia está diretamente ligada ao desenvolvimento.

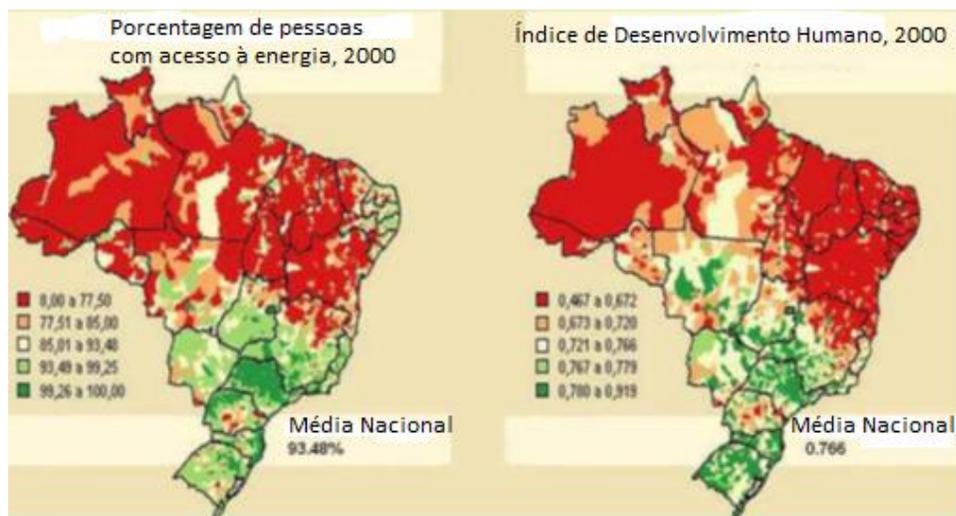


FIGURA 2.1 - Acesso à energia e índice de desenvolvimento humano
FONTE: XXIII ICOLD QUESTION 88 (2009)

A atenção aos aspectos sociais e ambientais das barragens e dos reservatórios deve ser uma preocupação prioritária, da mesma forma que a atenção à segurança é invariavelmente uma prioridade. A promoção de técnicas de engenharia deve ser orientada para o planejamento, o projeto, a operação e a manutenção de barragens seguras que garantam o desenvolvimento e a gestão sustentável dos recursos hídricos (CIGB-ICOLD, 2008).

Conforme já mencionado, as barragens são barreiras ou estruturas atravessando um rio, por exemplo, com o objetivo de represar e controle do fluxo de água para fins como: fornecimento de água para consumo, irrigação, geração de energia elétrica, permitir a navegação, controle de enchentes (secas) e recreação (CIGB-ICOLD, 2008). Os principais tipos de barragens são as construídas com concreto, chamadas de barragens de gravidade, em arco ou de contrafortes e as barragens em aterro que são construídas de terra ou de uma combinação de terra e rochas (enrocamento) (ICOLD - CIGB, 2008).

Segundo ICOLD – CIGB (2008), a água tem sido usada como fonte de energia desde a era romana. A princípio era usada para mover moinhos para vários processos mecânicos, tais como: moer cereais, cortar madeira ou movimentar a confecção de tecelagem. As turbinas elétricas foram criadas no século XIX quando a energia hidráulica começou a ser usada para gerar eletricidade. A hidroeletricidade consiste no uso da água em movimento para girar uma turbina conectada por um eixo a um

gerador. Pelo fato da água ser uma fonte de energia limpa, renovável e economicamente viável, a energia de fonte hidrelétrica é largamente utilizada para gerar desenvolvimento sustentável. De acordo com CIGB-ICOLD (2008), a hidroeletricidade responde por cerca de 20% de toda energia elétrica gerada. Estima-se que cerca de 25% do número total de grandes barragens tenha o único propósito de gerar energia elétrica.

A FIGURA 2.2 apresenta um esquema básico de funcionamento de uma usina hidrelétrica constituído de reservatório, barragem, casa de força com gerador, transformador e turbina, linhas de transmissão, portão de controle, comporta e saída de escoamento.

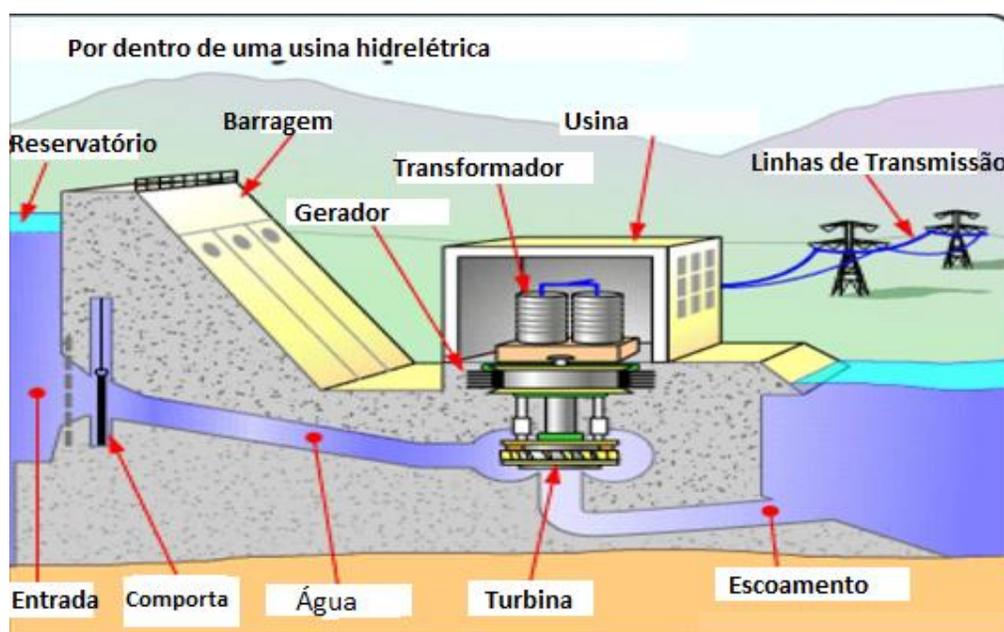


FIGURA 2.2 - Esquema básico de uma usina hidrelétrica
FONTE: XXIII ICOLD QUESTION 88 (2009, p. 281)

A capacidade instalada de geração de energia elétrica no Brasil, em janeiro de 2016, atingiu o montante de 141.684 MW, dos quais, 92.100 MW provenientes da fonte hidráulica. Em Janeiro de 2016, o Brasil contava com 4500 usinas de geração, das quais, 1220 de fonte hidráulica, conforme a TABELA 2.1 (MINISTÉRIO DE MINAS E ENERGIA, 2016).

TABELA 2.1 – Matriz da capacidade instalada de geração de energia elétrica no Brasil- Janeiro 2016 (MW)

FONTE	Nº USINAS	CAPACIDADE INSTALADA (MW)	(%)
<i>Hidráulica</i>	1.220	92.100	65,0
<i>Térmica</i>	2.916	41.595	29,4
Gás Natural	149	12.439	8,78
Biomassa	520	13.277	9,37
Petróleo	2.192	10.124	7,15
Carvão	22	3.612	2,6
Nuclear	2	1.990	1,40
Outros	31	153	0,10
<i>Eólica</i>	330	7.968	5,6
<i>Solar</i>	34	21	0,0
CAPACIDADE TOTAL BRASIL	4.500	141.684	100

FONTE: Ministério de Minas e Energia (2016, p. 15)

De modo geral, os órgãos relacionados às grandes barragens desejam assegurar que as barragens e suas estruturas associadas sejam seguras, econômicas e ambientalmente sustentáveis (ICOLD - CIGB, 2008).

A FIGURA 2.3 exibe a famosa Curva da Banheira representando a probabilidade de falhas em barragens. Ela sugere que há maior probabilidade de falhas no período do represamento da barragem, seguida de uma menor probabilidade de falhas no período de operação e o posterior aumento com o envelhecimento da barragem. Desta forma, é desejável prolongar o período operacional da barragem e ‘retardar’ seu envelhecimento através do contínuo investimento em ações para este fim.

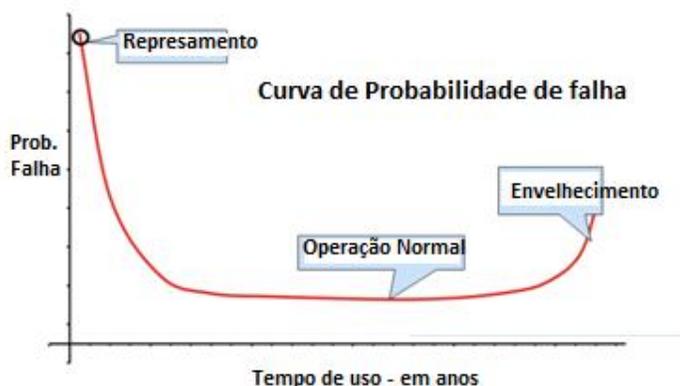


FIGURA 2.3 – Curva de probabilidade de falhas de uma barragem com o tempo
FONTE: XXIV ICOLD QUESTION 93 (2012, p. 39)

Recentemente, em maio de 2014, a usina de Itaipu completou 30 anos do início da geração e a liderança mundial em produção de energia limpa e renovável. Segundo Jornal Interno de Itaipu (2014), apesar de críticas na década de 80 quanto à necessidade de instalação de um potencial de produção de 12,6 mil megawatts (atualmente é de 14 mil MW) em tempos de economia estagnada, a decisão governamental binacional àquela época mostrou-se correta já que, em maio de 2014, Itaipu respondeu pela produção de 17% da energia elétrica do mercado brasileiro e 75% do mercado paraguaio. Na FIGURA 2.4 é exibida a representatividade das fontes de geração de energia no Brasil no ano de 2013.

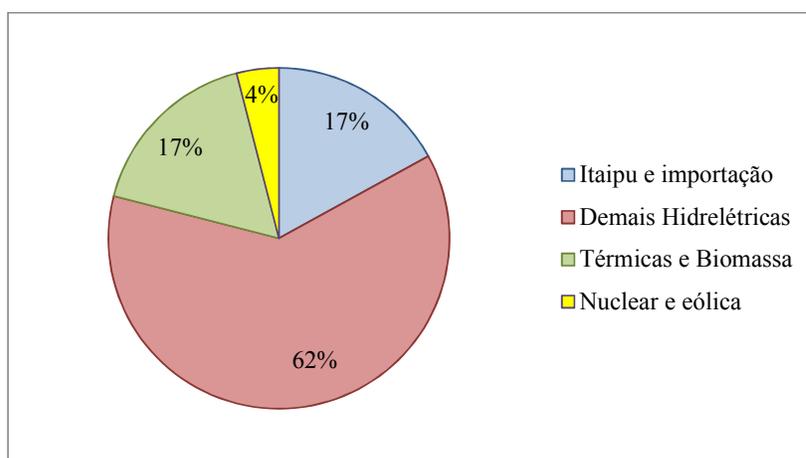


FIGURA 2.4 - Produção energética do Brasil em 2013 por fontes
FONTE: MINISTÉRIO DAS MINAS E ENERGIA (2013)

Em 30 anos de operação, Itaipu já gerou 2,16 bilhões de MWh, energia suficiente para abastecer o mundo inteiro por 38 dias e o Brasil por 4 anos e 8 meses (GAZETA DO POVO, 2014). Itaipu é atualmente a segunda maior hidrelétrica do mundo em potência instalada, 14 mil MW em 20 unidades geradoras, superada por

Três Gargantas na China que tem potência instalada de 22,4 mil MW. Mesmo assim, Itaipu ainda tem superado, na produção anual de energia, Três Gargantas que teve sua produção máxima de 98,11 MWh em 2012, em 2013 a usina estabeleceu pelo segundo ano consecutivo um novo recorde mundial ao produzir 98,6 milhões de MWh (G1, 2014). Em 2015, Itaipu voltou a liderar a produção anual de energia elétrica no mundo, produzindo 89,2 milhões de MWh (megawatt-hora) contra 87 MWh da usina de Três Gargantas (UOL, 2016).

2.1.2 Segurança e Instrumentação de barragens

O monitoramento do desempenho da barragem é um elemento crítico na produção de energia e manutenção de uma barragem segura. Os programas de monitoramentos consistem tipicamente de: fiscalização ou inspeção visual; instrumentação; coleta de dados; avaliação, apresentação e gestão dos dados; e decisões/respostas com base em resultados. Os programas de monitoramento e controle das barragens fornecem informações para avaliar o desempenho da barragem em relação ao objetivo do projeto e comportamento esperado, acompanhar mudanças que podem afetar o desempenho seguro da barragem, auxiliar nas investigações e avaliações de desempenho anormal ou degradação da barragem e determinar se é necessária manutenção ou ação corretiva (ASDSO - ASSOCIATION OF STATE DAM SAFETY OFFICIALS, 2015).

O uso da instrumentação cresceu muito com a construção de grandes barragens para atender a necessidade de informações adicionais do projeto e o interesse causado por falhas em barragens por todo o mundo, a fim de assegurar a segurança de barragens durante seu período operacional (ICOLD - BULLETIN 21, 1969).

Nos países em que o potencial hidrelétrico tem sido aproveitado extensivamente, este desenvolvimento teve início há mais de 50 anos atrás e muitas barragens já podem ser consideradas envelhecidas e é necessária a manutenção das condições de segurança que poderão evitar obras de reparação caras e quase sempre conflitantes com a produção de energia. O aprimoramento das barragens assegura não somente a operação e o desempenho seguro, mas também melhoria na sua eficiência (XXIII ICOLD QUESTION 88, 2009), (XXIII ICOLD QUESTION 90, 2009).

As barragens, por serem grandes obras de engenharia, estão sujeitas a grandes acidentes. A ruptura de uma barragem pode ocasionar extensos alagamentos a jusante e causar enorme impacto social, econômico e ambiental. Não basta uma construção segura, é necessário assegurar que sua operação também o seja. É necessário o acompanhamento e a avaliação permanente do comportamento da barragem. Isto se dá através de um processo denominado auscultação de barragem que consiste de inspeções visuais e por instrumentação (FUSARO, 2007).

Mesmo após longos períodos de operação sem problemas, o envelhecimento da estrutura de uma barragem deve se basear numa rotina de monitoramento cautelosa das condições da estrutura para fornecer avisos de emergência, minimizar a perda material e financeira e evitar vítimas humanas (KHEIFITS, 1995).

A segurança absoluta de uma barragem é impraticável do ponto de vista econômico e as barragens não estão livres de riscos, sendo necessário manter os riscos dentro de níveis aceitáveis. O monitoramento remoto é um conceito básico no suporte à segurança da barragem, pois qualquer comportamento anormal precedido da detecção de sinais de alerta em tempo, pode prevenir falhas e desastres (XXIII ICOLD QUESTION 91, 2009).

A Lei 12.334 (CASA CIVIL DA PRESIDÊNCIA DA REPÚBLICA, 2010) estabelece a política nacional de segurança de barragens do Brasil (PNSB), define em seu artigo 2º parágrafo III que “segurança de barragem é uma condição que visa manter a sua integridade estrutural e operacional e a preservação da vida, da saúde, da propriedade e do meio ambiente”, já no artigo 4º, destaca-se dentre outros fundamentos da PNSB, o inciso V: “a segurança de uma barragem influi diretamente na sua sustentabilidade e no alcance de seus potenciais efeitos sociais e ambientais”.

A mesma lei, em seu artigo 3º expõe os seguintes objetivos da PNSB:

- Garantir a observância de padrões de segurança de barragens de maneira a reduzir a possibilidade de acidente e suas consequências;
- Regulamentar as ações de segurança a serem adotadas nas fases de planejamento, projeto, construção, primeiro enchimento e primeiro vertimento, operação, desativação e de usos futuros de barragens em todo o território nacional;

- Promover o monitoramento e o acompanhamento das ações de segurança empregadas pelos responsáveis por barragens;
- Criar condições para que se amplie o universo de controle de barragens pelo poder público, com base na fiscalização, orientação e correção das ações de segurança;
- Coligir informações que subsidiem o gerenciamento da segurança de barragens pelos governos;
- Estabelecer conformidades de natureza técnica que permitam a avaliação da adequação aos parâmetros estabelecidos pelo poder público;
- Fomentar a cultura de segurança de barragens e gestão de riscos.

Deve ser dada atenção contínua a operação e a manutenção da segurança de uma barragem de usina. A ferramenta para isto é a instrumentação da barragem. O conceito básico da instrumentação é o monitoramento da infiltração, pressão e deslocamentos. Por conta das particularidades de cada usina hidrelétrica, não há procedimentos universais aplicáveis a todas as usinas no que se refere à avaliação da instrumentação. A avaliação e o julgamento das informações disponíveis por uma experiente equipe de engenharia são a melhor forma de contribuir para uma tomada de decisão e a escolha da melhor ação a ser executada (USACE, 1995).

A instrumentação é um conjunto de dispositivos instalados nas barragens ou perto delas para efetuar medições que podem ser usadas para avaliar o comportamento estrutural e os parâmetros de desempenho da estrutura (ICOLD - CIGB, 2008).

Dentre outros requisitos técnicos e administrativos, as barragens devem possuir instrumentação adequada para monitoramento do seu desempenho, o objetivo é permitir o controle da operação da barragem com segurança sob qualquer condição. As condições climáticas, afluência, nível de reservatório, vazão e níveis do rio a jusante devem ser monitoradas. A instrumentação instalada deve permitir a medição do comportamento estrutural e a condição física da barragem (USACE, 1995).

Até 1995, dos 17 maiores acidentes mundiais (relacionados a barragens, aterros e semelhantes) que custaram muitas vidas, 10 ocorreram durante os 5

primeiros anos de operação. Durante o período de 30 a 50 anos de operação livre de problemas na estrutura, mais do que uma geração de pessoal na operação adquire a confiança na segurança e com o início do envelhecimento da estrutura a equipe de manutenção operacional pode formar uma falsa percepção da segurança (KHEIFITS, 1995).

O comportamento da barragem tem que ser continuamente avaliado ao longo do tempo. Segundo Casa Civil da Presidência da República (2010), deve ser preconizada a realização periódica da segurança de barragens com o objetivo de verificar o estado geral de segurança da barragem. As mudanças no desempenho com o tempo, modificações corretivas de estruturas no decorrer de sua vida útil e danos inevitáveis requerem julgamento sobre o descarte, conserto ou substituição de instrumentos (USACE, 1995).

Um programa de instrumentação contribui muito para a segurança de uma barragem. Para a correta compreensão do comportamento de uma barragem, especialmente com o seu envelhecimento, é importante aplicar o conhecimento em quantidade e qualidade que for exigido. Deve haver pessoal qualificado, equipamentos de qualidade e informação adequada e rápida para avaliação (USACE, 1995). Em especial, esta é a proposta do CEASB – Centro Avançado de Estudo de Segurança de Barragem da usina hidrelétrica de Itaipu Binacional, localizada no Rio Iguaçu, entre o Brasil e o Paraguai.

Desde que a terra e as rochas são formações naturais, ao contrário de outros materiais da engenharia, tais como aço e concreto, estão mais sujeitas a incertezas, portanto, observações visuais e a instrumentação adequada fornecem informações importantes sobre a manutenção das premissas de um projeto e a avaliação do desempenho e a segurança de uma barragem de usina. Quando uma falha, uma condição severa, mudança visualmente observada ou uma infiltração tenha ocorrido em uma barragem, a instrumentação pode ser extremamente útil na determinação das causas do evento, bem como, quando ações corretivas são implantadas e se deseja avaliar o efeito do tratamento (USACE, 1995).

Outro benefício da análise dos dados de instrumentação é detectar um período de desempenho satisfatório. Isto pode servir para que em uma futura variação dos dados seja indicado algum potencial problema. Os dados também podem ser úteis na

previsão do desempenho sob condições máximas de operação como, por exemplo, no caso de enchentes. A análise do desempenho de uma barragem e dos dados de instrumentação gerados durante sua fase de operação pode ser usada no planejamento e na construção de novas barragens mais seguras e econômicas.

Tendo em vista a singularidade que uma barragem representa, a determinação da quantidade, tipo e localização de instrumentos em uma barragem exige a combinação de experiência, senso comum e intuição. A instrumentação deve ser concebida considerando as condições geotécnicas específicas da barragem, de sua fundação e do reservatório (USACE, 1995).

No planejamento de um sistema de instrumentação a previsão da magnitude da mudança é necessária para que o domínio de valores e a sensibilidade de um instrumento sejam selecionados. A estimativa do valor máximo (ou mínimo) possível ou de interesse leva a seleção do domínio do instrumento. Outro fator que não pode ser esquecido é o monitoramento de possíveis causas da variação dos instrumentos, quais sejam: nível do reservatório, quantidade de chuva, temperatura da água e do ambiente, pressão atmosférica e eventos sísmicos (USACE, 1995).

Os responsáveis pelo monitoramento da instrumentação devem ser capazes de avaliar se um instrumento está funcionando corretamente. Por exemplo, certa parte dos instrumentos pode ter aquisição automática de dados de leitura de instrumentação, mas nem por isso, leituras manuais devem ser descartadas, ao contrário, deve se confrontar ambas. Pode haver correlação entre instrumentos do mesmo tipo ou de tipos distintos localizados proximamente. Esta relação de dependência pode criar regras para avaliar o comportamento de alguns conjuntos de instrumentos (USACE, 1995).

Outra importante etapa de um sistema de instrumentação é o gerenciamento dos dados. O gerenciamento consiste na coleta, redução e processamento e apresentação. Em especial na apresentação, gráficos são necessários para: fornecer comparações visuais entre o comportamento real e o previsto, facilitar a detecção de erros na aquisição de dados, determinar tendências ou efeitos cíclicos, comparar o comportamento com outros instrumentos, prever o comportamento futuro, determinar eventual necessidade de manutenção da instrumentação. Os gráficos permitem que os dados sejam comparados facilmente com eventos que causam alterações nos

dados, tais como mudanças ambientais, bem como, avaliar o comportamento imprevisto e determinar a eficácia de uma ação corretiva.

A instrumentação de uma barragem fornece dados para determinar se a estrutura completa está funcionando conforme desejado e fornece uma supervisão contínua da estrutura para informar sobre algo que comprometa sua segurança (NEDUSHAN, 2002).

A análise dos dados consiste em sua avaliação considerando sua influência sob várias condições. Deve haver consciência da existência de dados inválidos e cálculos indevidos a fim de que não se incorra em interpretações incorretas. Uma análise adequada deve abordar dois aspectos básicos de monitoramento de segurança de barragens: o desempenho do sistema de instrumentação e o desempenho da estrutura ou característica que está sendo monitorada. Deve haver precaução sobre a precisão de gráficos de controle gerados automaticamente por computadores. Estes não devem ser utilizados e interpretados sem uma cuidadosa revisão de pessoal especializado (USACE, 1995).

A propósito, é importante citar a diferença entre acurado e preciso. Segundo Oakland (2008), acurado diz respeito à capacidade de atingir um valor alvo (a média, por exemplo), ou seja, tem a ver com tendenciosidade (vício) de um instrumento, enquanto que preciso relaciona-se com o grau de dispersão dos valores (a variância, por exemplo).

A interpretação dos dados deve ser realizada em tempo hábil, preferencialmente logo após a coleta dos dados. Programas de computador que geram gráficos de controle da leitura da instrumentação normalmente possuem limites de alerta baseados em critérios determinísticos, isto é, estabelecidos com base em critérios de projeto e que há poucos sistemas informatizados de alerta baseados em critérios estatísticos (FUSARO, 2007).

A propósito, os valores de controle para a leitura da instrumentação na barragem de Itaipu foram estabelecidos com base no método de elementos finitos (método numérico usado na determinação da solução de sistemas estruturais complexos, como barragens, modelos reduzidos estruturais e geomecânicos) e critérios de projeto (ITAIPU BINACIONAL, 1984).

Os valores obtidos de análises teóricas ou modelos físicos são baseados em certos parâmetros e hipóteses simplificadoras, salientando ainda que deva haver a permanente confrontação entre os valores estabelecidos no projeto com aqueles obtidos nas leituras, não impedindo que eventuais leituras ultrapassem os limites de projeto (ITAIPU BINACIONAL, 1999_a).

Uma definição para valores de controle é que:

Valores de controle são aqueles mais próximos da realidade física, considerando as influências térmicas ambientais atuantes sobre as estruturas e modelos reológicos mais realistas do concreto e da fundação. Os valores de controle são utilizados para o acompanhamento do desempenho das estruturas, por meio das leituras efetuadas “in situ”, e constituem sinais de alerta para situações anormais do comportamento estrutural. Estes valores, para o caso de Itaipu, já foram determinados anteriormente na fase de projeto e do enchimento do reservatório, não sendo, em muitos casos, mais aplicáveis na fase de operação (ITAIPU BINACIONAL, 1999_b, p. 1).

Faz-se necessário, portanto, o estudo para a reavaliação de valores de controle para a instrumentação de auscultação das estruturas da Central Hidrelétrica de Itaipu e suas fundações. A TABELA 2.2 apresenta a quantidade e localização da instrumentação em Itaipu.

TABELA 2.2 - Total de instrumentos de auscultação das estruturas civis de Itaipu

ESTRUTURA	LOCAL/TIPO DE INSTRUMENTO			TOTAL
	Fundação	Concreto	Geodesia	
Barragem de Terra –MD	35	--	14	49
Vertedouro	89	93	10	192
Barragem lateral direita	177	324	25	526
Barragem Principal	202	447	29	678
Estrutura de Desvio	91	235	5	331
Casa de Força	143	263	--	406
Barragem de enrocamento	60	--	38	98
Barragem de terra	63	--	40	103
TOTAL	860	1362	161	2383

FONTE: ITAIPU BINACIONAL (2002, p. 2)

De acordo com Fusaro (2007), as grandezas físicas avaliadas por instrumentos de auscultação de barragens geram séries temporais que podem ser tratadas por métodos estatísticos em substituição a modelos determinísticos ou analíticos que, neste caso, devido à geometria e ao contorno e à variabilidade nos parâmetros dos materiais de construção, tornam a implementação extremamente complexa não sendo possível eliminar todas as incertezas sobre os parâmetros geotécnicos. Neste trabalho não serão abordados métodos determinísticos.

Os valores da leitura de um instrumento de monitoramento de barragem possuem uma variabilidade inerente, associada ao próprio instrumento, a fatores humanos e a fatores de instalação e utilização (FUSARO, 2007).

A quantidade de esforço empregado na análise dos dados ainda é pequena e desproporcional em relação ao esforço empregado na instrumentação e coleta dos dados na barragem (NEDUSHAN, 2002).

O monitoramento de barragens não tem acompanhado os avanços recentes da análise estatística e há necessidade de desenvolver novas ferramentas de análise para contribuir com a equipe de segurança de barragem na avaliação do comportamento da barragem (DIBIAGIO¹, 2000, *apud* NEDUSHAN, 2002).

Segundo USACE (1995), a análise técnica dos dados deve considerar as seguintes técnicas quando se analisa os dados:

- Comparar dados atuais com o conjunto de dados mais recentes para detectar anomalias, alterações no comportamento de curto prazo e mau funcionamento de instrumentos.
- Comparar os dados atuais com o desempenho histórico durante um período significativo de tempo para determinar a consistência do desempenho da característica monitorada pelo instrumento.
- Comparar as tendências de comportamento ao longo do tempo com tendências previstas durante o planejamento ou com valores relativos a fatores de

¹ DIBIAGIO, E. Monitoring of dams and their foundations. Pequim: 20th. International Commission on Large Dams, q. 78, General Report. 2000.

segurança calculados. Neste ponto, o comportamento histórico de uma estrutura torna-se a base do comportamento futuro e o desempenho previsto durante o projeto torna-se menos relevante.

- Comparar, quando possível, leituras da instrumentação com sistemas complementares para confirmar ou negar uma mudança física implícita.
- Utilizar análises estatísticas para avaliar o desempenho dos instrumentos. Sistemas automatizados (ou manuais) podem gerar uma grande quantidade de dados que favorecem o cálculo de estatísticas como média e desvio padrão. Isto também é útil para a determinação da frequência de calibração.

Uma das mais importantes considerações associadas ao planejamento e desempenho da instrumentação é uma crítica avaliação dos dados. Além disso, a enorme quantidade de dados deve ser resumida, erros de leitura devem ser rejeitados e a interpretação deve ser criteriosa. A apresentação de gráficos deve ser simples e de fácil entendimento (ICOLD - BULLETIN 21, 1969).

Segundo ICOLD² (1989, *apud* FUSARO, 2007) é relatado que a análise dos dados de instrumentação requer ciência de que:

- Existe um *time-lag* (intervalo entre dois fenômenos relacionados, assim como causa e efeito) entre um fenômeno físico e a resposta dada por um instrumento relacionado;
- Escalas distorcidas em gráficos podem levar a interpretações e conclusões equivocadas;
- Conclusões devem se basear em tendências estabelecidas ao longo de um período de tempo razoável das observações;
- Correlações com diferentes tipos de dados devem ser estabelecidas de forma que garantam confiabilidade aos processos de monitoramento;
- As limitações inerentes de cada instrumento devem ser previamente conhecidas a fim de evitar tentativas sem sucesso de avaliação de dados cuja magnitude está dentro da margem de erro do instrumento;

² ICOLD. **Bulletin 68**. International Comitee of Large Dams. Paris. 1989.

- A ocorrência de dados de instrumentação em desacordo com os valores previstos não implica necessariamente a existência de um problema, bem como, não se pode afirmar com certeza que dados dentro dos valores previstos impliquem na inexistência de problemas.

Conforme Nedushan (2002), há dois modelos tradicionais de monitoramento de barragens: o modelo estatístico e o modelo determinístico. No modelo estatístico é estudada a correlação entre fatores ambientais de entrada como nível do reservatório, temperatura ambiente, nível de chuvas e as saídas de leituras de instrumentação como infiltração, deslocamento, deslizamento, tensão, entre outras. Esta correlação é estimada com base na análise do histórico de dados. Modelos estatísticos buscam responder a questão: a barragem está se comportando como no passado? Já os modelos determinísticos buscam modelar a barragem para prever seu comportamento em longo prazo em situações extremas de nível do reservatório, temperatura, eventos sísmicos, entre outras. Modelos determinísticos buscam responder a questão: a barragem irá se comportar adequadamente a determinada condição?

Algumas consequências da análise dos dados da instrumentação apontadas por USACE (1995) são:

- Determinar quando testar, calibrar ou descartar um instrumento.
- Determinar se o cronograma de observações deve ser alterado.
- Reavaliar que áreas do projeto exigem atenção prioritária.
- Determinar a necessidade de um estudo mais aprofundado.
- Confirmar ou refutar estudos anteriores.
- Preparar os dados processados para apresentação formal a equipe de segurança ou gerencia.

Com o passar do tempo, os instrumentos devem passar por manutenção para que continuem confiáveis.

Em um dos trabalhos pioneiros interessados no gerenciamento dos dados de instrumentação de barragens e revisão da literatura até então disponível, Nedushan (2002) afirma que a instrumentação tem o objetivo de validar modelos numéricos, fornecer ideias sobre o comportamento da barragem, detectar anomalias e permitir respostas rápidas, seja na forma de reparos, manejo do nível do reservatório ou, até

mesmo, evacuação. Salienta ainda que dentre as tarefas que devem ser executadas estão a identificação de instrumentos com defeitos, remoção de dados discrepantes, interpretação dos dados, ajustamento de modelos e a gestão de alarmes de detecção de mudanças estatisticamente significativas nas respostas da barragem.

Dentre as principais anomalias que a instrumentação pode detectar estão: a infiltração (*piping*) através dos maciços de terra e suas fundações, eventuais trincas com infiltrações no concreto, movimentos bruscos entre os blocos da barragem, subpressões excessivas na cabeça dos blocos a montante, ou a jusante da Casa de Força, movimentos excessivos na região da calha do vertedouro (ITAIPU BINACIONAL, 2002).

Em 1999 foi feita uma reavaliação dos valores de controle da instrumentação na usina de Itaipu. Essa calibração dos modelos numéricos executados em Itaipu considerou num determinado período de condição normal de utilização, a ação de carregamentos oriundos de cargas hidrostáticas (empuxo da água que age sobre um bloco e subpressões no pé da barragem que provém de fissuras na região) e de variações sazonais da temperatura sobre os blocos que possuem maior instrumentação e comparou seus resultados com as leituras dos instrumentos de auscultação. Este estudo buscava reavaliar os valores de controle e estabelecer um modelo estatístico entre grandezas *causa* como variação térmica, empuxo causado pela variação do nível do reservatório e grandezas *efeito* como deslocamentos, infiltrações, deformações, subpressões. Em resumo, buscou calibrar os modelos determinísticos e propor modelos estatísticos (ITAIPU BINACIONAL, 1999_c).

Alguns instrumentos de auscultação instalados nas estruturas de concreto de Itaipu e sua respectiva função estão destacados no QUADRO 2.2.

QUADRO 2.2 - Instrumentos nas estruturas de concreto de Itaipu

INSTRUMENTO	FUNÇÃO
Pêndulo Direto	Medir deslocamento da crista da barragem na direção do fluxo e normal a ele em relação a sua base
Termômetro	Medir estado térmico da barragem
Tensômetro	Tensão no concreto (recalque e distensão)

FONTE: ITAIPU BINACIONAL (1984; 1999_a)

Outros instrumentos instalados nas estruturas de concreto são: deformímetro, medidores de junta, marcos de referência e células de empuxo.

Instrumentos de auscultação instalados nas fundações da barragem de Itaipu e sua respectiva função são destacados no QUADRO 2.3.

QUADRO 2.3 – Instrumentos nas fundações da barragem de Itaipu

INSTRUMENTO	FUNÇÃO
Pêndulo invertido	Medir deslocamento da base da barragem em relação a pontos profundos da fundação.
Piezômetro <i>standpipe</i>	Medir subpressão no contato concreto-rocha e em níveis mais permeáveis do maciço basáltico da fundação.
Extensômetro múltiplo de haste	Deslocamentos (deformações) verticais (recalques) e inclinados (na direção do furo de sondagem).
Roseta de extensômetro múltiplo	Deslocamentos horizontais da base da barragem em relação a pontos profundos na fundação.
Base de alongâmetro entre blocos	Recalques diferenciais entre blocos e deslocamentos diferenciais na direção horizontal.

FONTE: ITAIPU BINACIONAL (1984)

Na FIGURA 2.5 é apresentado um esquema resumido do posicionamento de pêndulos diretos e invertidos em um bloco da barragem de Itaipu.

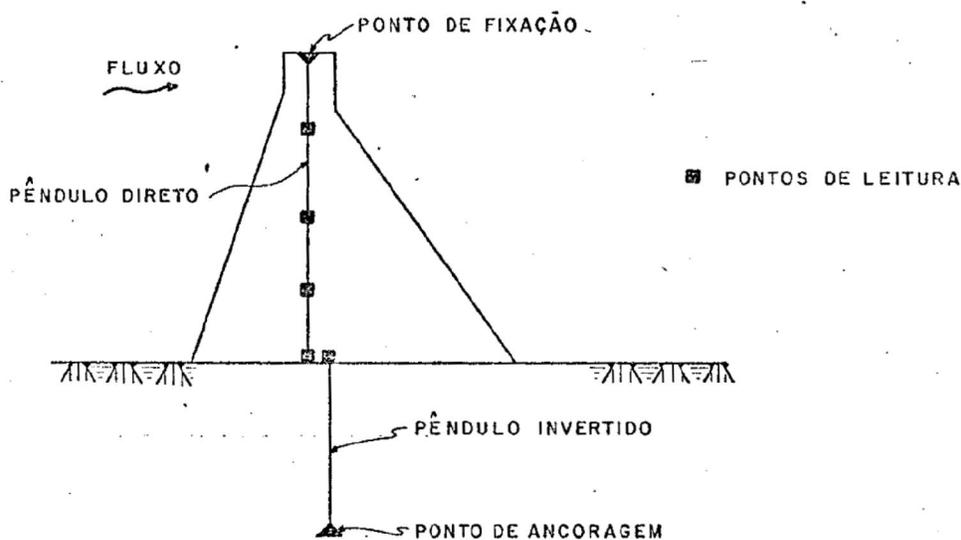


FIGURA 2.5 - Esquema de funcionamento de pêndulos diretos e invertidos em Itaipu
FONTE: ITAIPU BINACIONAL (1984)

Na FIGURA 2.6 é apresentado o esquema de funcionamento de um extensômetro múltiplo constituído de três hastes em Itaipu.

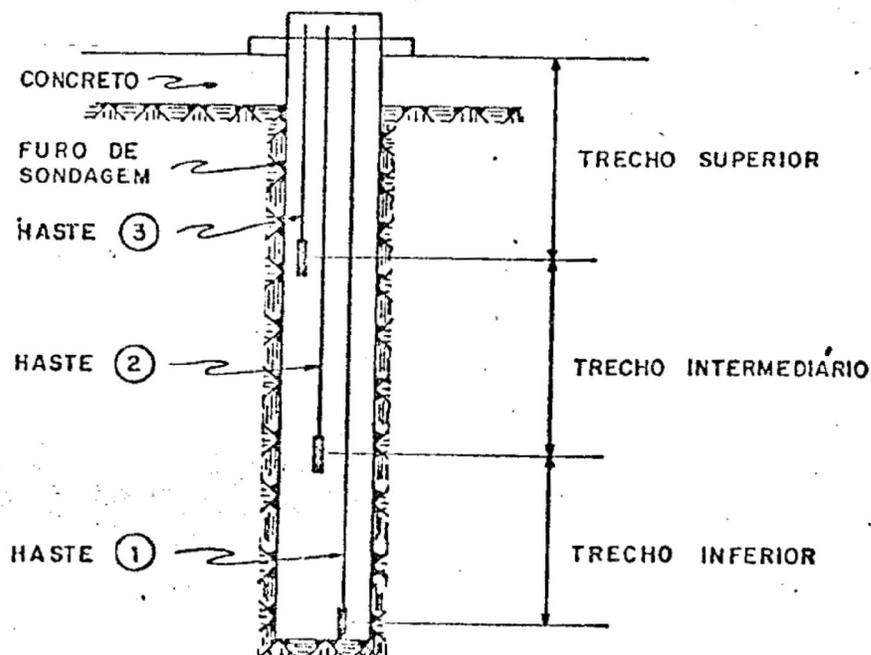


FIGURA 2.6 - Esquema de um extensômetro múltiplo em Itaipu
FONTE: ITAIPU BINACIONAL (1984)

2.1.3 Controle Estatístico de Qualidade

2.1.3.1 Introdução

Pode ser especulado que a segunda guerra mundial foi vencida pelo controle de qualidade e o uso da estatística moderna (RYAN, 2011, p. 6).

Os métodos estatísticos e outras técnicas de resolução de problemas podem melhorar a qualidade de produtos tais como: bens manufaturados, serviços ou outros produzidos por qualquer área de uma companhia. A qualidade é um conceito subjetivo que está relacionado às características desejáveis de um produto ou serviço. Fato é que cada vez mais é um fator importante na decisão de um consumidor na escolha de um produto ou serviço, seja este consumidor um indivíduo, uma empresa ou um governo (MONTGOMERY, 2013).

A proposta do controle estatístico de qualidade é que um produto fornecido a um cliente atendendo suas especificações atenda, também, a satisfação do consumidor. A inspeção de todos os produtos é cara e ineficiente, mas as consequências do envio de produtos não conformes podem ser prejudiciais em termos da insatisfação do cliente. O controle estatístico de qualidade é o processo de inspeção na linha de produção de amostras dos produtos em produção ou inspeção de amostras de lotes para aceitação desses lotes, com o objetivo de garantir probabilisticamente um nível de qualidade especificado (NATIONAL INSTITUTE OF STANDARDS TECHNOLOGY, 2014).

Uma definição moderna de qualidade é: “*qualidade é inversamente proporcional à variabilidade*”, onde a variabilidade seja não desejada ou danosa, ou seja, quanto menor a variabilidade, maior é a qualidade. Uma companhia automotiva americana realizou nos anos 80 um estudo comparativo da transmissão de um modelo seu e de uma concorrente japonesa. Os resultados mostraram que os custos da companhia japonesa eram menores, sobretudo por causa dos termos de garantia e custos de reparos. Seleccionadas amostras aleatórias de ambas para medir características de qualidade das transmissões, observou-se que a distribuição das características críticas da transmissão americana ocupava uma amplitude das especificações bem maior que a japonesa (FIGURA 2.7), ambas com mesma e distribuição normal, mas com diferentes desvios padrões, do que se concluiu que havia muito menos variabilidade nas características de qualidade nas transmissões fabricadas no Japão e que havia maior chance de produção fora do padrão na transmissão fabricada nos Estados Unidos e esta era a causa dos custos americanos serem superiores (MONTGOMERY, 2013).

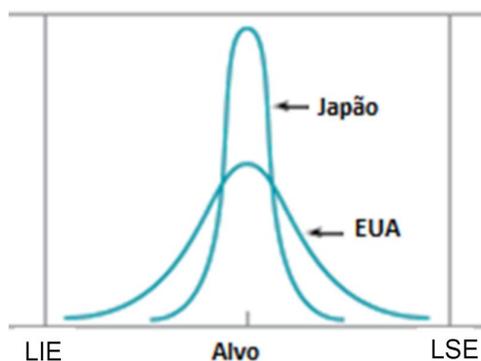


FIGURA 2.7 – Distribuições das dimensões críticas para as transmissões
FONTE: MONTGOMERY (2013, p. 3)

A prevenção de falhas é o principal objetivo da melhoria de um processo que uma equipe de gerenciamento busca. O sistema idealizado para alcançar a melhoria continua é um ciclo interminável, chamado de Ciclo de Deming, conforme a FIGURA 2.8, que fornece a estratégia em que as técnicas de controle estatístico de qualidade são úteis e identificar as etapas da melhoria (OAKLAND, 2008).

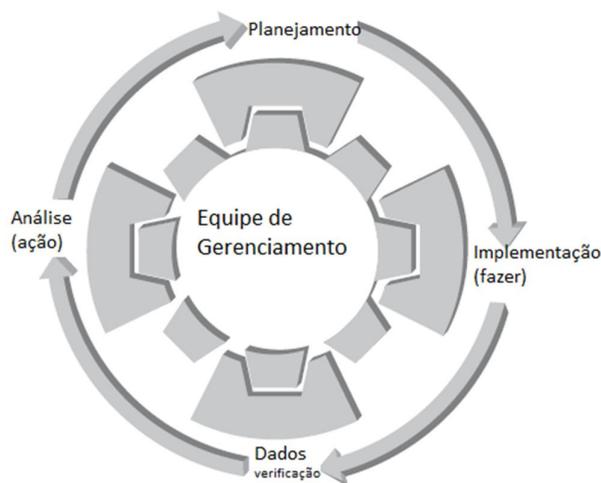


FIGURA 2.8 – Ciclo de Deming
FONTE: OAKLAND (2008, p. 347)

A variabilidade é inerente a qualquer característica de qualidade de um produto ou serviço. Os métodos estatísticos desempenham importante papel no esforço da melhoria de qualidade de um produto, pois a variabilidade só pode ser descrita em termos estatísticos. A avaliação de uma característica de qualidade seja esta física, sensorial, ou de orientação temporal, se dá em função de suas especificações. A especificação é um intervalo de tolerância contendo o valor desejado (nominal, alvo) de uma característica de qualidade, o maior valor permitido de uma característica de qualidade é o limite superior de especificação (LSE) e o menor é o limite inferior de especificação (LIE). As especificações geralmente resultam do planejamento de engenharia de um produto (MONTGOMERY, 2013).

O início formal do controle estatístico de qualidade é considerado em 1924, quando Walter A. Shewhart introduziu o conceito estatístico de gráfico de controle na *Bell Telephone Laboratories*. Após a segunda guerra mundial houve grande expansão e aceitação do controle estatístico de qualidade nas indústrias de manufatura. No

início dos anos 80 as companhias ocidentais, ao descobrirem que competidores japoneses utilizavam planejamento de experimentos na avaliação e desenvolvimento de processos, melhoria da confiabilidade e desempenho dos produtos, tolerância de componentes, despertaram maior interesse em experimentos estatísticos (MONTGOMERY, 2013). Shewhart entendeu que monitorar a variabilidade estatisticamente numa fábrica poderia conduzir os processos e produtos a melhores níveis de qualidade e menor variabilidade das características avaliadas. Aumentar a qualidade significa reduzir custos e aumentar o resultado financeiro de uma empresa (SAMOBYL, 2009).

Técnicas estatísticas são necessárias para determinar se alguma variação anormal do que está sendo monitorado tenha ocorrido, para determinar mudanças nos parâmetros do processo e identificar fatores que influenciam as características do processo (RYAN, 2011). O controle estatístico de qualidade é uma estratégia para a redução da variabilidade, na causa da maioria dos problemas de qualidade, na forma de fazer as coisas, nos materiais, na atitude das pessoas, nos equipamentos e no seu uso, nas práticas de manutenção, enfim, em tudo (OAKLAND, 2008).

Uma das principais técnicas de controle estatístico do processo é o gráfico (ou carta) de controle. Um típico gráfico de controle (FIGURA 2.9) consiste em plotar as médias das medidas de uma característica de qualidade em função do tempo ou da amostra. O gráfico possui uma linha central que representa o valor alvo da característica de qualidade caso não houvesse variabilidade e de duas outras linhas, o limite inferior e superior de controle que são determinados com base estatística (MONTGOMERY, 2013).



FIGURA 2.9 - Um típico gráfico de controle
FONTE: Adaptado de MONTGOMERY (2013)

O gráfico de controle é uma técnica de monitoramento do processo amplamente utilizada. Se alguma média amostral cair fora do intervalo dos limites de controle, isto é sinal de que a variabilidade está excessiva e investigações adicionais devem ser realizadas para reduzir a variabilidade e melhorar a qualidade (MONTGOMERY, 2013).

Para Oakland (2008), conforme mostrado na FIGURA 2.10, um gráfico de controle tem três zonas nas quais a ação depende em qual zona um resultado é obtido. Na primeira zona (estável) a ação é continuar normalmente, não fazer nada, pois se está numa zona com somente causas comuns de variação. A segunda zona (de alerta) informa que é preciso cautela e busca de informações adicionais, pois o processo pode estar com causas especiais de variação. Na terceira (de ação) zona é preciso agir, investigar ou até ajustar o processo, pois causas especiais de variação estão presentes.

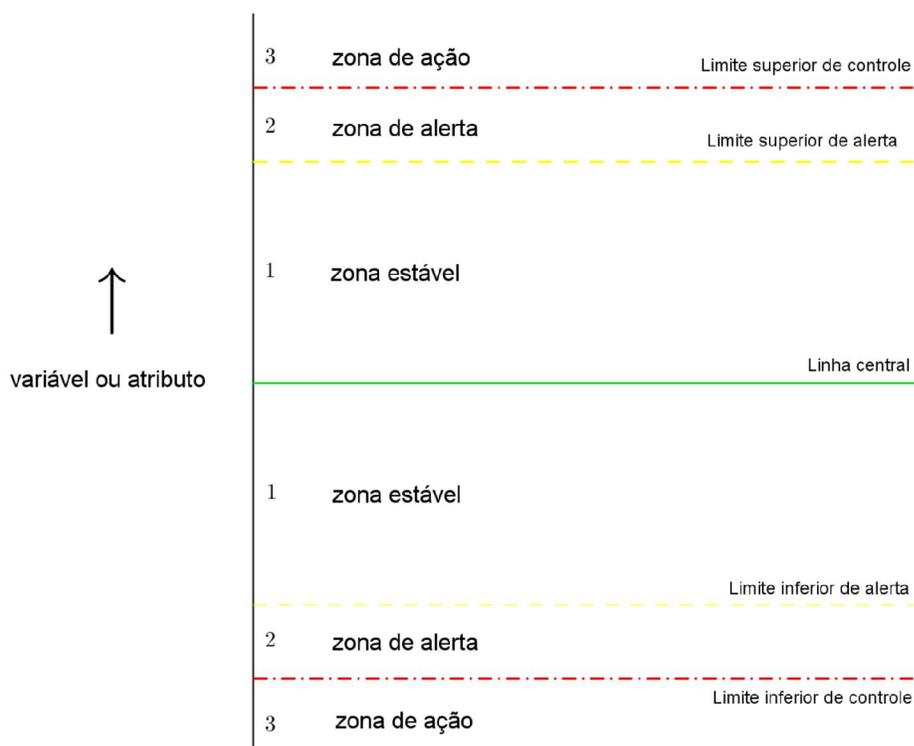


FIGURA 2.10 - Esquema de um gráfico de controle
FONTE: Adaptado de OAKLAND (2008)

Conforme já mencionado, qualquer processo de produção possui uma variabilidade inerente, independentemente de seu planejamento ou manutenção, essa

variabilidade natural ou inevitável é chamada de “*ruído de fundo*” é o efeito acumulativo de várias pequenas causas no sistema de controle de qualidade. A variabilidade natural constitui o denominado sistema estável de causas aleatórias. Um processo que opera somente com causas aleatórias de variação é dito estar sob controle estatístico (MONTGOMERY, 2013, p. 96).

Desta forma, o principal objetivo do Controle Estatístico de Processo é identificar o mais breve possível a presença de variabilidade anormal (causas especiais de variação) e eliminá-las. Para tal, o gráfico de controle é uma técnica muito útil, no sentido de que é eficaz na identificação de variabilidade anormal do processo (MONTGOMERY, 2013, p. 97).

Quando qualquer tipo de violação da estabilidade de um processo for identificado, os gráficos de controle não devem ser usados e as causas especiais devem ser investigadas e eliminadas e os limites de controle recalculados após a exclusão de pontos fora de controle oriundos de causas especiais. Este procedimento melhora a estimativa devida somente a causas comuns. Num primeiro momento, a maioria dos processos industriais não está sob controle e as causas especiais de fora de controle têm que ser encontradas e corrigidas (OAKLAND, 2008).

Um modelo geral para um gráfico de controle de uma estatística amostral w que mede alguma característica de qualidade de interesse com média μ_w e desvio padrão σ_w terá como limite superior de controle (LSC), valor da linha central (LC) e limite inferior de controle (LIC) os valores

$$\begin{aligned} LSC &= \mu_w + L\sigma_w, \\ LC &= \mu_w, \\ LIC &= \mu_w - L\sigma_w, \end{aligned} \tag{2.1.1}$$

onde L é a distância dos limites de controle à linha central expressa em unidades de desvio padrão. Como estes princípios foram estabelecidos inicialmente por Walter S. Shewhart, tais gráficos de controle são chamados de gráficos de controle de Shewhart.

Um dos principais benefícios dos gráficos de controle é a manutenção de um bom histórico, seja para aplicações administrativas ou de controle do processo (ou ambos). Um bom histórico de dados é essencial para tornar real o controle de qualidade. Sua ação na redução da variabilidade vai desde o seu efeito psicológico

sobre a equipe de trabalho até a efetiva identificação de causas especiais de variação (RYAN, 2011).

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

$$\begin{aligned}\alpha &= P(\text{erro tipo I}) = P(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}), \\ &= P(\text{rejeitar o estado de controle dado que o processo está sob controle}), \\ &= P(\text{falso alarme}),\end{aligned}\quad (2.1.2)$$

α é chamado de nível de significância do teste e

$$\begin{aligned}\beta &= P(\text{erro tipo II}) = P(\text{aceitar } H_0 | H_0 \text{ é falsa}), \\ &= P(\text{aceitar o estado de controle dado que o processo está fora de controle}), \\ &= P(\text{falso positivo}).\end{aligned}\quad (2.1.3)$$

Chama-se de poder do teste ao valor:

$$\begin{aligned}1 - \beta &= P(\text{rejeitar } H_0 | H_0 \text{ é falsa}) \\ &= P(\text{rejeitar o estado de controle} | \text{processo não está sob controle}).\end{aligned}\quad (2.1.4)$$

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 L e calcular o valor de α ou vice-versa (MONTGOMERY, 2013, p. 102).

Por exemplo, se a característica de qualidade de interesse μ_w tem distribuição normal pode se calcular que

$$P(\mu_w \in (\mu_w - 3\sigma_w, \mu_w + 3\sigma_w)) = 0,9973 = 1 - \alpha. \quad (2.1.5)$$

Em gráficos de controle, 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. Já, cometer um erro tipo II significa rejeitar que o processo está fora de controle quando, na verdade, realmente está fora de controle. O erro tipo II é normalmente avaliado em gráficos de controle pelo comprimento médio da sequência (ARL ou ARL_1) necessário para identificar que o processo está fora de 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 sequência (CMS) tradução do inglês de *average run length (ARL)* que é o número médio de observações que se espera tomar 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). Quando o processo não está sob controle é desejável diminuir o tempo necessário para que o gráfico de controle detecte um alarme verdadeiro mais rapidamente (PINTO, 2009), (YANG e SHEU, 2006).

No gráfico de controle de Shewhart para um processo não autocorrelacionado o ARL é calculado por

$$ARL = \frac{1}{P(\text{uma observação apareça fora de controle})} = \frac{1}{\alpha} \quad , \quad (2.1.6)$$

em que α é a probabilidade de que uma observação não pertença ao intervalo da região de controle. Por exemplo, para o gráfico de controle da média amostral (\bar{x}) com limites de $\pm 3\sigma$ onde $\alpha = 0,0027$ o valor do $ARL = 370$, significa que mesmo que o processo esteja sob controle, *em média*, a cada 370 observações espera-se que uma observação esteja fora do intervalo ($\bar{x} - 3\sigma, \bar{x} + 3\sigma$). Denota-se o ARL do gráfico de controle de \bar{x} para um processo sob controle por ARL_0 que é calculado por

$$ARL_0 = \frac{1}{\alpha} \quad (2.1.7)$$

e o ARL do gráfico de controle de \bar{x} para um processo fora de controle por ARL_1 que é calculado por

$$ARL_1 = \frac{1}{1 - \beta} \quad . \quad (2.1.8)$$

Intuitivamente, para observações independentes, o número de observações necessárias até que uma caia fora de um limite de controle é uma variável aleatória geométrica com parâmetro p cuja esperança matemática é $\frac{1}{p}$, justamente o ARL (MONTGOMERY, 2013, p. 104).

Quanto maior o ARL_0 e menor o ARL_1 mais eficaz é um gráfico de controle, pois consegue perceber rapidamente a falta de controle quando este sofre alguma modificação e demorar a acusar falta de controle quando este não sofre qualquer alteração (PINTO, 2009).

Dentre as técnicas de monitoramento de processos, os gráficos de controle amostrais da média \bar{x} , da variância S e da amplitude R estão entre os mais úteis. A teoria estatística subjacente à construção de limites em gráficos de controle de uma característica de qualidade pode presumir que esta seja normalmente distribuída ou não. 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 (MONTGOMERY, 2013).

Os limites de especificação medem a tolerância permitida da variabilidade de uma característica de qualidade de interesse. A tolerância é calculada pelo Engenheiro na concepção do processo ou produto, antes de sua 'fabricação'. Já os valores de controle são calculados com base na observação de valores práticos e não teóricos. Em suma, a tolerância mede o que devia ser, enquanto os limites de controle medem o que de fato é (SAMOHYL, 2009). Não há qualquer relação matemática ou estatística entre os limites de controle e os limites de especificação. Normalmente, é claro, os limites de controle estão contidos nos limites de especificação (MONTGOMERY, 2013, p. 138).

É importante citar que normalmente 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 para assegurar que o processo está sob controle. Isto quer dizer que o processo está operando próximo de um valor alvo aceitável com alguma variação natural e sem causas de preocupação (CHAKRABORTI, 2011). 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. Dados

históricos são utilizados para calcular os limites de controle, em seguida os dados são comparados com os limites iniciais estabelecidos e pontos que estejam fora dos limites são investigados e, algumas vezes, descartados, caso em que os limites podem ser recalculados para estabelecer a fase I e a posterior aplicação na fase II de dados em tempo real do monitoramento de um processo (NATIONAL INSTITUTE OF STANDARDS TECHNOLOGY, 2014).

Quando o tamanho da amostra é $n = 1$, isto é, forma-se um gráfico de controle de Shewhart para medidas individuais provenientes de uma distribuição normal. Isto ocorre em diversas situações práticas por razões como: a tecnologia de inspeção e medição automática é empregada em toda unidade do processo, a taxa de produção é lenta não permitindo amostras maiores que um, há pouca variabilidade dentro de uma amostra, dentre outras. Este caso será de especial interesse neste trabalho, tendo em vista futuras aplicações.

Uma desvantagem dos gráficos de controle de Shewhart é que estes são relativamente insensíveis a pequenas mudanças no processo, da ordem de até $1,5\sigma$. Duas alternativas para lidar com isso são o gráfico de controle da soma cumulativa (CUSUM – *CUmulative SUM* em inglês) e da média móvel exponencial ponderada MMEP (*Exponentially Weight Moving Average* – EWMA em inglês) (MONTGOMERY, 2013).

O uso de gráficos de Shewhart ou CUSUM 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 (RYAN, 2011, p. 139).

2.1.3.2 Gráficos de controle multivariados

Nas subseções 2.1.3.2, 2.1.3.3 e 2.1.3.4 são discutidas a importância do tratamento de dados de forma multivariada, as estratégias para lidar com gráficos multivariados e alguns gráficos de controle multivariados mais utilizados.

Gráficos de controle multivariados foram introduzidos por H. Hotelling (HOTELLING, 1947), na época da segunda guerra mundial e, a partir do

desenvolvimento dos computadores, que viabilizaram sua aplicação, tiveram grande desenvolvimento.

Avaliar o desempenho e controlar a qualidade de processos industriais complexos é de importância cada vez maior nas indústrias químicas, de construção civil e, em geral. Algumas razões para isto são legislações cada vez mais rigorosas com base na segurança de processos, questão ambiental, aumento da concorrência global. Tarefas associadas implicam acompanhamento *on-line* em Engenharia Civil, Mecânica, Automobilística, Aeroespacial, Elétrica e Eletrônica. Consequentemente, tem aumentado a preocupação com o monitoramento da exatidão e da precisão da instrumentação (KRUGER e XIE, 2012).

A alta quantidade de características monitoradas e o recebimento *on-line* de dados de monitoramento fez crescer no meio industrial e acadêmico, nos últimos 20 a 30 anos, do 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 como um grupo e não separadamente (MASON e YOUNG, 2002).

Segundo Fusaro (2007), a utilização da análise de um conjunto de instrumentos para a análise mais global de uma seção ou região de uma barragem para estabelecer alertas sobre o comportamento de um maciço de terra ou enrocamento, permitiria que situações de risco pudessem ser mais consistentemente caracterizadas desta forma do que a análise isolada de um instrumento.

Para Mason e Young (2002), o controle do processo multivariado é uma metodologia baseada em gráficos de controle que são usados para monitorar a estabilidade do processo multivariado. A estabilidade é atingida quando as médias, variâncias e covariâncias das variáveis do processo se mantêm estáveis sobre amostras das observações.

O exemplo a seguir ilustra a necessidade da avaliação multivariada de um processo. Suponha que as leituras de temperatura e pressão de 4 pontos (dados) disponíveis de um processo sejam: $A = (178; 76)$, $B = (180; 80)$, $C = (170; 70)$, $D =$

(172; 74). Na FIGURA 2.11 a seguir são apresentados os quatro pontos, a média amostral do processo $M = (\bar{x}_1, \bar{x}_2) = (175; 75)$, uma reta de ajuste dos pontos (mínimos quadrados) e dois 'círculos' centrados na média.

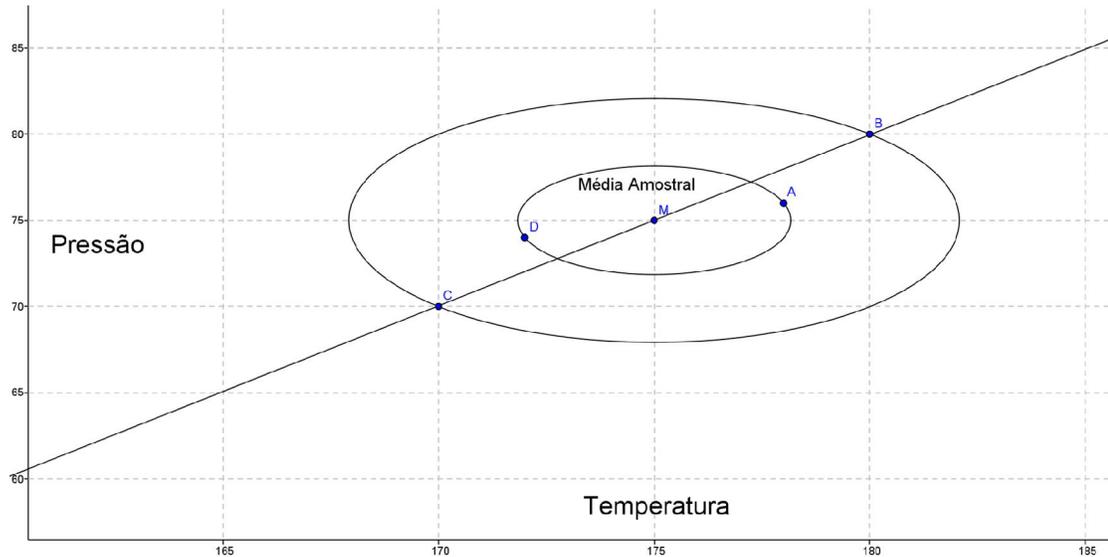


FIGURA 2.11 - Diferença entre distância estatística e distância Euclidiana
FONTE: Adaptado de Mason e Young (2002, p. 7)

Na FIGURA 2.11 parece que a temperatura tem mais variação do que a pressão, contudo, isso pode ser devido à diferença na escala das variáveis no gráfico, mas neste caso é, de fato, verdade. Quando se ignora a correlação entre as variáveis (o que está claro neste caso que na medida em que a temperatura cresce a pressão também cresce) e simplesmente se analisa a distância (Euclidiana) de cada ponto à média amostral se obtém, respectivamente, os valores (3,16;7,07;7,07;3,16).

A distância estatística é concebida para levar em conta a diferença na variação entre as variáveis. Sejam

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2, i = 1,2, \quad (2.1.9)$$

a variância amostral da variável x_i ($i = 1$ (temperatura) e $i = 2$ (pressão)) onde \bar{x}_i é a média amostral e

$$s_{12} = \frac{1}{n-1} \sum_{j=1}^n (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2), \quad (2.1.10)$$

a covariância entre as duas variáveis e $n = 4$ são os quatro dados do exemplo, então a distância estatística (DE) quadrática de um ponto (x_1, x_2) à média (\bar{x}_1, \bar{x}_2) pode ser calculada usando

$$(DE)^2 = \frac{1}{1 - r^2} \left[\left(\frac{x_1 - \bar{x}_1}{s_1} \right)^2 - 2r \left(\frac{x_1 - \bar{x}_1}{s_1} \right) \left(\frac{x_2 - \bar{x}_2}{s_2} \right) + \left(\frac{x_2 - \bar{x}_2}{s_2} \right)^2 \right], \quad (2.1.11)$$

onde

$$r = \frac{s_{12}}{s_1 s_2} \quad (2.1.12)$$

é o coeficiente de correlação amostral entre as variáveis. Pelo fato da distância estatística estar relacionada com a equação de uma elipse esta é, às vezes, chamada de distância elíptica, distância de *Mahalanobis*, distância T^2 de Hotelling ou simplesmente distância T^2 . Em particular, se as variáveis não são correlacionadas, $r = 0$ e DE reduz-se a distância de um ponto ao vetor de médias, ponderada pelas variâncias.

Calculando a distância estatística quadrática de cada um dos quatro pontos dados à média obtém-se que todos distam 1,5 da média. Isto visualmente não parece ser razoável, visto que, pela FIGURA 2.11, os pontos A e D estão mais próximos da média do que os pontos B e C . Entretanto, quando existe diferença na variância ($s_1 \neq s_2$) e correlação entre as variáveis ($s_{12} \neq 0$) Montgomery (2013), Mason e Young (2002) e Ryan (2011) salientam que a distância estatística é o instrumento mais adequado nos procedimentos de controle multivariados. Especificamente neste exemplo os pontos B e C poderiam indicar sinais de 'fora de controle' quando o processo está sob controle (erro tipo I), uma vez que sua distância Euclidiana da média é maior que dos pontos A e D , entretanto, a distância estatística de todos à média, de fato, é a mesma. Desta forma, deseja-se enfatizar que os procedimentos de controle estatístico multivariado diferem da mera análise da distância Euclidiana.

Suponha agora que duas variáveis z_1 e z_2 sejam independentes e possuam distribuição normal padrão. Considerando que o processo está sob controle, se uma amostra cair dentro dos limites de controle três sigmas de cada variável, uma observação pode estar simultaneamente dentro dos limites de controle univariados das duas variáveis, porém, esta observação aparenta ter comportamento incomum

comparado aos demais. Isto é observada na FIGURA 2.12, no ponto marcado com o símbolo \otimes (onde LSC e LIC representam o limite superior e inferior de controle).

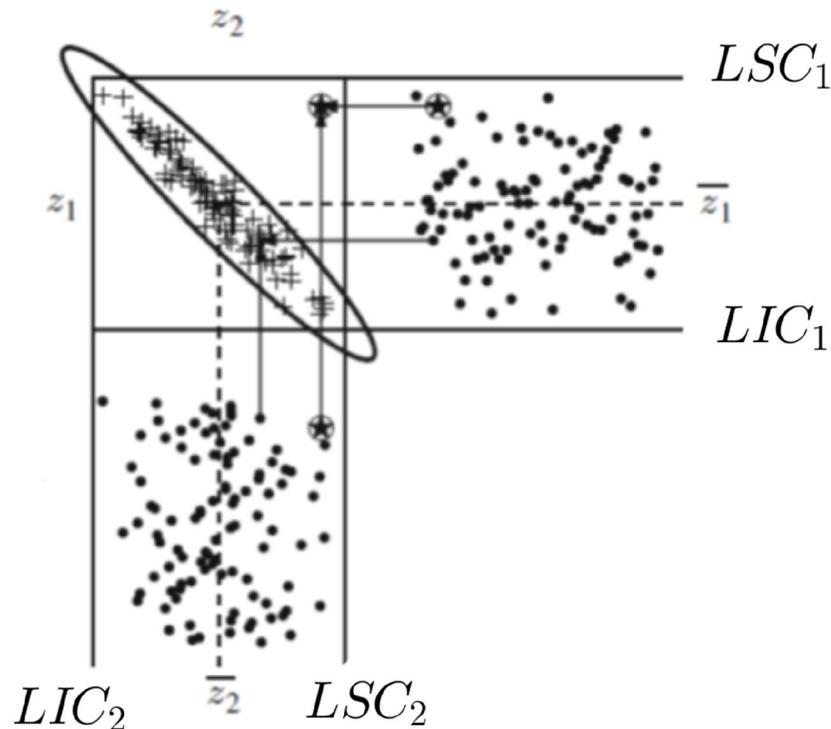


FIGURA 2.12 - Esquema gráfico de duas variáveis altamente correlacionadas
 FONTE: KRUGER e XIE (2012, p. 20)

De acordo com Montgomery (2013) e Ryan (2011), neste caso, a probabilidade conjunta de ambas as variáveis excederem seus limites de controle três sigma quando ambas estão sob controle simultaneamente é $(0,0027)^2 = 7,29 \times 10^{-6} \ll 0,0027$. Ou seja, a probabilidade de erro tipo I, se os gráficos de controle univariados e independentes são considerados, não mantém o mesmo valor do que os gráficos de controle individuais.

Por outro lado, a probabilidade de que ambas as variáveis de uma dada observação estejam dentro dos limites de controle quando o processo está sob controle é $(0,9973)^2 = 0,9946$. Segundo Montgomery (2013), o uso de gráficos independentes gera uma distorção no monitoramento simultâneo das variáveis. No exemplo anterior, a probabilidade de que ao menos um dos gráficos de controle sinalize uma condição de fora de controle assumindo os limites três sigma é de $1 - (1 - 0,0027)^2 = 0,00539$. Em geral, quando p variáveis são monitoradas

independentemente, se a probabilidade de erro tipo I, α , é igual para todas as variáveis, então a probabilidade de erro tipo I conjunta é

$$\alpha' = 1 - (1 - \alpha)^p. \quad (2.1.13)$$

Enquanto que a probabilidade de que todas as variáveis caiam dentro de seus limites de controle é $(1 - \alpha)^p$. Estas expressões não são válidas quando as variáveis são correlacionadas ou não são independentes. Caso os gráficos individuais sejam ajustadas em função de um valor α' predeterminado para a taxa de falsos alarmes multivariada, então o valor de α deve ser

$$\alpha = 1 - (1 - \alpha')^{1/p}. \quad (2.1.14)$$

Para enfatizar, se $\alpha = 0,0027$ é a probabilidade de erro tipo I para cada uma de um conjunto de $p = 19$ variáveis independentes e normalmente distribuídas, então a probabilidade α' de erro tipo I do processo multivariado conjunto das $p = 19$ variáveis, que significa a probabilidade de ao menos uma variável estar fora dos seus limites de controle é

$$\alpha' = 1 - (1 - 0,0027)^{19} \cong 0,05. \quad (2.1.15)$$

Há uma forte necessidade do uso de gráficos de controle multivariados quando as características de qualidade são correlacionadas. Assuma que duas variáveis sejam altamente e positivamente correlacionadas, como na FIGURA 2.13. Se os pontos começam a cair fora da elipse de controle, analisar as variáveis independentemente pode conduzir a erros. Recomenda-se o uso de um gráfico de controle multivariado ao invés do uso de um gráfico de controle para cada variável (RYAN, 2011).

Note ainda na FIGURA 2.13 que há uma região em que um ponto cai fora da elipse (ponto *C*), mas ainda dentro do retângulo, neste caso seria cometido erro tipo II (indicar que o processo está sob controle quando na verdade não está – falso verdadeiro). Por outro lado há uma região em que um ponto está dentro da elipse, mas fora do retângulo (ponto *D*), neste caso seria cometido erro tipo I (indicar que o processo está fora de controle quando na verdade está sob controle – falso alarme). Nos pontos *A* e *B* as respostas fornecidas pela região retangular e pela elipse de controle seriam as mesmas, controle e fora de controle, respectivamente.

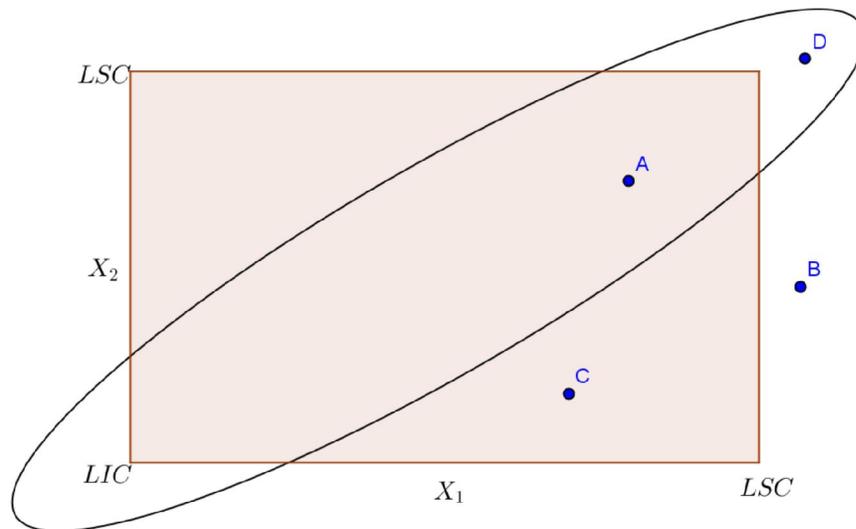


FIGURA 2.13 - Elipse de controle bivariada e região de controle retangular de 2 gráficos de Shewhart
FONTE: O autor

Os gráficos de controle multivariados podem ser usados quando as amostras sejam de observações individuais ou não. As desvantagens dos gráficos de controle multivariados em relação aos univariados é que sua compreensão não é tão fácil, quando um sinal de fora de controle é recebido é mais difícil identificar qual ou quais variáveis contribuíram para isso, se houve mudança na média, na variância ou na estrutura de correlação (RYAN, 2011).

2.1.3.3 Gráfico de controle T^2

Considere que p variáveis estejam sendo controladas, assumindo que $\mathbf{X}' = [X_1, X_2, \dots, X_p] \sim N_p(\boldsymbol{\mu}, \Sigma)$, ou seja, que \mathbf{X}' tenha distribuição normal multivariada e que se deseja controlar a média $\boldsymbol{\mu}$ do processo, bem como 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) significaria que o processo está fora de controle. Neste caso, Montgomery (2013) e Johnson e Wichern (2007) 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 (RYAN, 2011).

A estatística T^2 introduzida por Hotelling pode ser considerada uma generalização da estatística t que possui distribuição t -de Student (ou Gosset) onde

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \quad (2.1.16)$$

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

$$t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n} = n(\bar{x} - \mu_0)[s^2]^{-1}(\bar{x} - \mu_0). \quad (2.1.17)$$

Logo, no caso p dimensional, segue que

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'[S]^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0), \quad (2.1.18)$$

onde S^{-1} é a inversa da matriz de covariância (os elementos da diagonal principal são as variâncias e os elementos fora da diagonal principal são as covariâncias entre as p variáveis).

No caso de m amostras, de um processo sob controle, extraídas para avaliação serem de tamanho n , se p variáveis estiverem sendo avaliadas, segundo Montgomery (2013), como na prática quase sempre $\boldsymbol{\mu}$ e Σ são desconhecidas e são estimadas, respectivamente, pelos estimadores não viesados $\bar{\bar{\mathbf{X}}}' = (\bar{\bar{x}}_1, \dots, \bar{\bar{x}}_p)$ (média amostral) e a matriz de covariância amostral, que é simétrica positiva definida, dada por

$$S = \begin{bmatrix} \bar{S}_1^2 & \dots & \bar{S}_{1p} \\ \vdots & \ddots & \vdots \\ & \dots & \bar{S}_p^2 \end{bmatrix}, \quad (2.1.19)$$

onde:

$$\bar{\bar{x}}_j = \frac{1}{m} \sum_{k=1}^m \bar{x}_{jk} \quad j = 1, \dots, p, \quad (2.1.20)$$

representa a média sobre todas as m amostras da j -ésima variável,

$$\bar{S}_j^2 = \frac{1}{m} \sum_{k=1}^m S_{jk}^2 \quad j = 1, \dots, p, \quad (2.1.21)$$

representa a média sobre todas as m amostras das variâncias,

$$\bar{S}_{jh} = \frac{1}{m} \sum_{k=1}^m S_{jhk}^2 \quad j = 1, \dots, p; j \neq h, \quad (2.1.22)$$

representa a média sobre todas as m amostras das covariâncias entre as variáveis,

$$S_{jhk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ijk} - \bar{x}_{jk})(x_{ihk} - \bar{x}_{hk}) \quad k = 1, \dots, m; j \neq h, \quad (2.1.23)$$

representa a covariância entre as variáveis j e h na amostra k ,

$$S_{jk}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ijk} - \bar{x}_{jk})^2 \quad k = 1, \dots, m; j = 1, \dots, p, \quad (2.1.24)$$

representa a variância da variável j na k -ésima amostra

$$\bar{x}_{jk} = \frac{1}{n} \sum_{i=1}^n x_{ijk} \quad k = 1, \dots, m; j = 1, \dots, p, \quad (2.1.25)$$

representa a média da j -ésima variável na k -ésima amostra e x_{ijk} é a i -ésima observação da j -ésima variável na k -ésima amostra, então a estatística torna-se

$$T^2 = n(\mathbf{X} - \bar{\mathbf{X}})' [\mathbf{S}]^{-1} (\mathbf{X} - \bar{\mathbf{X}}). \quad (2.1.26)$$

Esta estatística é chamada de T^2 de Hotelling ou somente T^2 .

Segundo Ryan (2011) e Alt³ (1985, *apud* Montgomery, 2013) existem duas fases (ou estágios) no processo de controle estatístico de qualidade. Na primeira fase, chamada de retrospectiva, os dados de um processo servem para estabelecer o controle a partir de amostras preliminares. Nesta fase, os limites de controle são estabelecidos e testados para os dados disponíveis que devem estar sob níveis estaticamente considerados sob controle. Na segunda fase, chamadas de perspectiva, os limites de controle são estabelecidos a partir dos mesmos dados preliminares e servem para o monitoramento de dados futuros.

Os limites de controle da fase I e da fase II são mais próximos na medida em que o tamanho da amostra aumenta. Isto serve também para o caso univariado. Entretanto, Lowry e Montgomery (1995) alertam que no caso multivariado deve haver

³ ALT, F. B. Multivariate Quality Control. **Encyclopedia of Statistical Sciences v.6**, 1985.

maior cautela com a fixação dos mesmos limites de controle em ambas as fases. Portanto, Ryan (2011) e Montgomery (2013) afirmam que o limite superior de controle (LSC), nas condições de normalidade multivariada do gráfico de controle T^2 de Hotelling na fase I pode ser determinado por

$$LSC = \frac{p(m-1)(n-1)}{m(n-1)-p+1} F_{\alpha,p,mn-m-p+1}. \quad (2.1.27)$$

O limite inferior de controle é 0, por S se tratar de uma matriz positiva definida. Na fase II tem-se que

$$LSC = \frac{p(m+1)(n-1)}{m(n-1)-p+1} F_{\alpha,p,mn-m-p+1}. \quad (2.1.28)$$

Quando o número de observações de uma amostra é $n = 1$, situação que Montgomery (2013) relata ocorrer com frequência na indústria química e de processamento, bem como no caso de leitura de instrumentação de monitoramento de barragens, que será de interesse neste trabalho, a estatística T^2 de Hotelling conforme Montgomery (2013), Johnson e Wichern (2007) e Ryan (2011) é dada por $T^2 = (\mathbf{X} - \bar{\mathbf{X}})' [\mathbf{S}]^{-1} (\mathbf{X} - \bar{\mathbf{X}})$, ou simplesmente,

$$T^2 = (\mathbf{X} - \bar{\mathbf{X}})' [\mathbf{S}]^{-1} (\mathbf{X} - \bar{\mathbf{X}}) \quad (2.1.29)$$

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

$$LSC = \frac{p(m+1)(m-1)}{m(m-p)} F_{\alpha,p,m-p}. \quad (2.1.30)$$

Seguindo a indicação de Tracy, Young e Mason⁴ (1992, *apud* Montgomery 2013), o limite superior de controle (LSC) da fase I, no caso de observações individuais ($n = 1$) é recomendável calcular se baseando numa distribuição beta, neste caso

$$LSC = \frac{(m-1)^2}{m} \beta_{\alpha, \frac{p}{2}, \frac{m-p-1}{2}} \quad (2.1.31)$$

e o elipsoide p dimensional de $100(1 - \alpha)\%$ de previsão de uma futura observação é dado por todos os vetores X satisfazendo

⁴ TRACY, N. D.; YOUNG, J. C.; MASON, R. L. Multivariate Control Charts for individual observations. **Journal of quality technology**, v. 24, n. 2, p. 88-95, April 1992.

$$(X - \bar{X})'[S]^{-1}(X - \bar{X}) \leq LSC, \quad (2.1.32)$$

onde m é a quantidade de amostras (tempo). Em outras palavras, o valor dado na equação (2.1.31) é o limite superior de controle (LSC) para o gráfico de controle multivariado p dimensional da estatística T^2 de Hotelling da fase I. Nas equações, (2.1.30) e (2.1.31) tem-se, respectivamente, que $F_{\alpha,p,m-p}$ é o $1 - \alpha$ percentil da distribuição F de Snedecor com p e $m - p$ graus de liberdade e $\beta_{\alpha, \frac{p}{2}, \frac{m-p-1}{2}}$ é o $1 - \alpha$ percentil da distribuição β (beta) com $p/2$ e $(m - p - 1)/2$ graus de liberdade.

Quando se trata com dados normais multivariados de uma observação X que consiste de um vetor de dimensão p , de m amostras de tamanho $n = 1$, o comportamento de observações independentes pode ser descrito por uma função de probabilidade com parâmetros (μ, Σ) conhecidos ou desconhecidos e estimados por (\bar{X}, S) . Assume-se a existência de um conjunto histórico de dados, coletados sobre condições estáveis, quando um processo é considerado sob controle que pode ser usado para obter estimativas dos parâmetros desconhecidos.

Um resumo das propriedades da distribuição de probabilidade da estatística $T^2 = (X - \mu)'[\Sigma]^{-1}(X - \mu)$ é dado QUADRO 2.4, assumindo que a distribuição de probabilidade subjacente é normal multivariada.

QUADRO 2.4 – Valores e distribuição da estatística T^2 conforme os parâmetros e a fase

PARÂMETROS	Relação entre X e \bar{X} e entre X e S	VALOR E DISTRIBUIÇÃO DE T^2	APLICAÇÃO
Conhecidos (μ, Σ)	--	$T^2 = (X - \mu)'[\Sigma]^{-1}(X - \mu) \sim \chi_p^2$	Fases I e II
Desconhecidos e estimados por (\bar{X}, S)	Dependentes (X é usado para calcular \bar{X} e S)	$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X})$ $\sim \left[\frac{p(m+1)(m-1)}{m(m-p)} \right] F_{p,m-p}$	Fase I
Desconhecidos e estimados por (\bar{X}, S)	Independentes (X não é usado para calcular \bar{X} e S)	$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X}) \sim \left[\frac{(m-1)^2}{m} \right] \beta_{\frac{p}{2}, \frac{m-p-1}{2}}$	Fase II

FONTE: MASON e YOUNG (2002)

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

$$S_1 = \frac{1}{m} \sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})'. \quad (2.1.33)$$

É sugerido por Holmes e Mergen (1993) o uso do estimador da matriz de covariância a partir de pares de diferenças sucessivas, dado por

$$V_i = X_{i+1} - X_i \quad i = 1, \dots, m - 1. \quad (2.1.34)$$

Os vetores V_i podem ser arranjados em uma matriz

$$V = \begin{bmatrix} V_1' \\ \vdots \\ V_{m-1}' \end{bmatrix}, \quad (2.1.35)$$

de modo que a matriz

$$S_2 = \frac{1}{2(m-1)} V'V \quad (2.1.36)$$

é um estimador da matriz de covariância. Em Montgomery (2013, p. 332) se exhibe um exemplo em que o uso do estimador S_2 é mais capaz de detectar mudança no vetor média mesmo quando as matrizes de covariância estimadas das duas formas, não são estatisticamente diferentes, já Ryan (2011) aponta que esta forma de determinação do estimador da matriz de covariância é o análogo multivariado do método das médias móveis para gráficos univariados.

Para dados normais multivariados, a matriz de covariância usual S_1 é um estimador não viciado e de máxima verossimilhança, a função de probabilidade que descreve S_1 é conhecida e isto é importante na determinação da distribuição da estatística T^2 (MASON e YOUNG, 2002).

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 Runger, Alt e Montgomery (1996),

$$d_j = T^2 - T_{(j)}^2, \quad (2.1.37)$$

onde $T_{(j)}^2$ 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.

2.1.3.4 Gráfico de controle para dados autocorrelacionados

Os denominados ‘modelos de Shewhart’ de um processo sob controle de uma variável x no instante t têm a forma $x_t = \mu + \varepsilon_t$, onde se presume que x tem média μ e desvio padrão σ (fixos e desconhecidos) e os ε_t são independentes e normalmente distribuídos com média 0 e desvio padrão σ . A condição de fora de controle representa uma mudança em μ , em σ ou em ambos (MONTGOMERY, 2013, p. 289).

A hipótese da independência (autocorrelação) das observações em gráficos de controle é fundamental, pois gráficos de controle convencionais não funcionam bem na ausência da independência, através da geração de muitos falsos alarmes (MONTGOMERY, 2013, p. 289). Experimentalmente, o fenômeno da autocorrelação se reflete, por exemplo, na obtenção de observações relativamente grandes seguirem observações acima da média e da mesma forma para as relativamente pequenas.

O exame de padrões e tendências em gráficos univariados e a constatação de correlações semelhantes aos pares de variáveis podem ser indicadores de autocorrelação. Padrões especiais no gráfico da estatística T^2 podem indicar a presença de autocorrelação. Ignorar a autocorrelação pode enfraquecer o controle dado por T^2 . O efeito de uma variável autocorrelacionada sobre T^2 pode ser confundido com a sua dependência no tempo e relações de uma variável autocorrelacionada com outras variáveis podem ser mascaradas (MASON e YOUNG, 2002).

A medida analítica da correlação de uma série temporal (observações orientadas no tempo) é estimada pela função de autocorrelação amostral de defasagem (*lag*) k

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}. \quad (2.1.38)$$

Quando a autocorrelação é significativa, a estrutura de correlação dos dados pode ser apropriadamente modelada por uma série temporal que pode ser usada para ‘remover’ a autocorrelação e aplicar gráficos de controle aos resíduos (MONTGOMERY, 2013, p. 291). O primeiro modelo sugerido é um modelo autorregressivo de primeira ordem $AR(1)$ em que

$$x_t = \mu + \phi x_{t-1} + \varepsilon_t, \quad (2.1.39)$$

onde μ e $-1 < \phi < 1$ são desconhecidos e os ξ_t são normal e independentemente distribuídos com desvio padrão σ . Pode-se mostrar que

$$E(x_t) = \frac{\mu}{1 - \phi}, \quad (2.1.40)$$

$$V(x_t) = \frac{\sigma^2}{1 - \phi^2} \Rightarrow \sigma_{x_t} = \frac{\sigma}{\sqrt{1 - \phi^2}} \quad (2.1.41)$$

e observações que distam k períodos como $(x_t$ e $x_{t-k})$ tem coeficiente de correlação ϕ^k , de modo que a função de autocorrelação decai exponencialmente, uma vez que $|\phi| < 1$. Os parâmetros μ e ϕ são estimados pelo método dos mínimos quadrados, de modo a minimizar a soma dos erros ε_t . Finalmente, seja \hat{x}_t a estimativa de x_t obtida dos parâmetros μ e ϕ , os resíduos no instante t , definidos por

$$e_t = x_t - \hat{x}_t, \quad (2.1.42)$$

são normais e independentes com média 0 e variância constante, de sorte que se podem aplicar gráficos de controle convencionais à sequência de resíduos. Pontos fora de controle deste gráfico indicam que μ ou ϕ se alterou e que a variável original x_t está fora de controle.

A maioria dos procedimentos de controle multivariados exige que os vetores de observações sejam não correlacionados ao longo do tempo. A ausência desta hipótese pode enfraquecer a eficácia do processo de controle global. Se a autocorrelação não for detectada ou ignorada, pode criar problemas nos processos de controle que não se ajustam a ela. O principal problema é similar ao que ocorre quando se usam procedimentos univariados em um processo multivariado no qual a relação entre as variáveis é ignorada. Se a dependência temporal de dados autocorrelacionados não for removida, os efeitos de uma variável podem ser confundidos com os efeitos de outras variáveis correlacionadas (MASON e YOUNG, 2002).

Procedimentos de controle para dados autocorrelacionados no caso univariado, geralmente buscam modelar a autocorrelação e analisar o gráfico dos resíduos (diferença entre os valores observados e ajustados do modelo de autocorrelação), como os modelos ARIMA (seção 2.1.5). Com base em certas hipóteses, os erros residuais, isto é, quando o efeito da dependência no tempo é removido, pode se mostrar que são independentes e normalmente distribuídos.

No caso multivariado, gráficos das variáveis individualmente ao longo do tempo podem ser influenciadas por outras variáveis correlacionadas, provocando interpretações equivocadas. No caso de uma variável cíclica no tempo, conforme a FIGURA 2.14 que exibe o gráfico da série temporal de um piezômetro em Itaipu, medido em *msnm* – metros sobre nível do mar, a suspeita de alguma forma de autocorrelação pode não ser verdadeira e sim devido a alguma tendência sazonal. Mesmo assim, gráficos de cada variável, não deixam de ser úteis (MASON e YOUNG, 2002).

Os métodos de controle de qualidade multivariado tradicionais assumem que os vetores de observações são independentes no tempo. Mastrangelo, Runger e Montgomery (1996) mostram que quando há autocorrelação no processo, o gráfico T^2 de Hotelling pode atenuar a autocorrelação no sentido de que em algumas situações, técnicas multivariadas baseadas em T^2 são menos sensíveis a autocorrelação do que técnicas univariadas.

Variações cíclicas ou sazonais ao longo do tempo são assumidas serem baseadas em causas sistemáticas (FIGURA 2.14), isto é, a variação não ocorre aleatoriamente, mas é reflexo da influencia de variáveis 'escondidas'. A autocorrelação produz algum tipo de padrão sistemático ao longo do tempo nas observações de uma variável. Caso não corrigido, estes padrões são transformados em padrões não aleatórios no gráfico da estatística T^2 . A autocorrelação provoca o aumento da variação do valor de T^2 , o que pode dificultar a sensibilidade desta estatística em detectar mudanças pequenas e consistentes. Variações não aleatórias desta forma são explicáveis, mas podem e devem ser removidas (MASON e YOUNG, 2002).

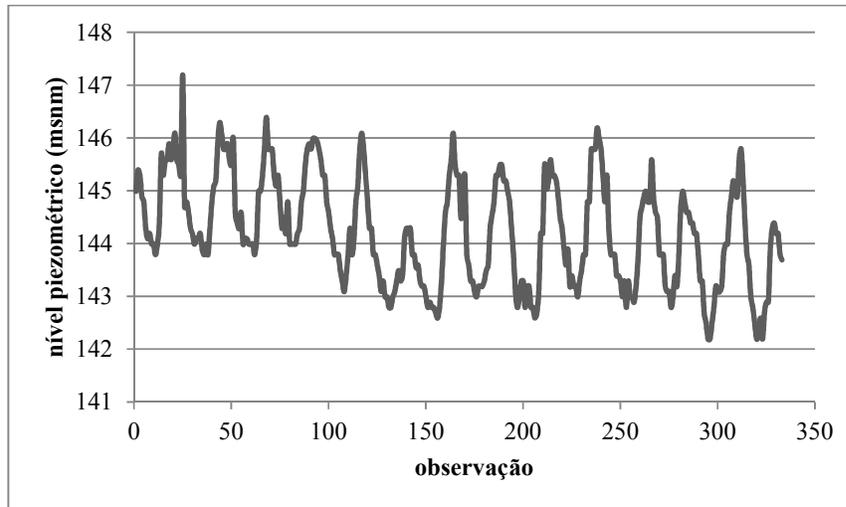


FIGURA 2.14 - Variável cíclica no tempo
FONTE: O autor

O modelo de Shewhart de gráfico de controle para uma variável nada mais é do que um caso especial dos modelos ARIMA em que os coeficientes *AR* e *MA* são todos nulos (MONTGOMERY, 2013, p. 296).

2.1.4 Análise de Componentes Principais (ACP)

Esta subseção faz uma revisão sobre a análise de componentes principais, uma alternativa para 'reduzir' a quantidade de dados em problemas multivariados, sem 'grande perda' de informação.

Técnicas de redução de dados são baseadas no princípio da construção de conjuntos de variáveis latentes 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 sob controle estatístico. Técnicas de redução e análise de dados podem ser divididas em: monobloco (análise de componentes principais (ACP), análise discriminante linear ou de Fisher, dentre outras) de duplo bloco (análise de correlação canônica, Regressão de Posto reduzido e Mínimos quadrados parciais, dentre outras) (KRUGER e XIE, 2012, p. 28).

Análise de componentes principais (ACP) é uma técnica analítica multivariada de dados na qual variáveis correlacionadas são transformadas em um conjunto em um número menor de variáveis não correlacionadas que são combinações lineares das variáveis originais. O método remonta a Karl Pearson em 1901 e a um artigo de Harold Hotelling em 1933. O grande impulso do método deu-se com o desenvolvimento dos computadores para aplicação em problemas de grande porte (JACKSON, 1991, p. xv).

De acordo com Nedushan (2002, p. 24), 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.

A análise de componentes principais (ACP) e os mínimos quadrados parciais (PLS) são dois métodos estatísticos multivariados que projetam os dados de variáveis altamente correlacionadas e ruidosas em subespaços de baixa dimensão que contêm informação relevante sobre o processo. Quando o conjunto de variáveis inclui variáveis de entrada e de saída, PLS é uma forma de modelar a estrutura latente e reduzir a dimensão do problema (MASTRANGELO, RUNGER e MONTGOMERY, 1996).

Para Johnson e Wichern (2007, p. 430) a análise de componentes principais é concebida para explicar a estrutura da matriz de covariância de um conjunto de variáveis através de algumas poucas combinações lineares destas variáveis com o objetivo de reduzir e interpretar os dados. A ideia é que de um número de p variáveis, a variabilidade total do sistema possa ser explicada, sem grande perda de informação, por um conjunto de k componentes principais que são combinações lineares das variáveis originais. Geometricamente, as componentes principais representam as coordenadas de um novo sistema obtido pela rotação do sistema, obtido das variáveis iniciais, que representam direções de máxima variabilidade e fornecem uma forma mais simplificada de descrever a estrutura de covariância.

A ACP pode ser extremamente útil em aplicações de controle de qualidade, pois ela permite transformar um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas que pode ser mais fácil de serem monitoradas via gráficos de controle. Sua popularidade provavelmente deriva de sua habilidade, em muitas ocasiões, de representar uma situação multivariada em um espaço de dimensão muito mais reduzida. Isto se refere a parcimônia: a capacidade de economia de meios para um dado fim. A parcimônia é algo desejável, embora nem sempre possível (JACKSON, 1991, p. 19).

A ACP é uma técnica estatística que pode ajudar a detectar a existência da ‘quase’ singularidade, determinar subgrupos de variáveis que são altamente correlacionadas e estimar a dimensionalidade de um sistema e, em alguns casos, dar ideia sobre a natureza destas dimensões. Uma orientação para identificar uma quase singularidade é o número de condição da matriz de covariância, que é a razão entre o maior autovalor e o menor, se este número for maior do que 30, indica que há uma severa colinearidade entre as variáveis. Neste caso, não é aconselhável usar a estatística T^2 , pois esta pode ser aumentada e são sugeridas alternativas com a exclusão de variáveis ‘redundantes’ ou usar um número de componentes principais com pouca perda de informação (MASON e YOUNG, 2002).

A colinearidade refere-se a correlação entre variáveis independentes. Uma hipótese da regressão múltipla é que não exista colinearidade perfeita entre as variáveis de entrada. Se existir, então as estimativas de mínimos quadrados não podem ser encontradas (PANKRATZ, 1991).

Um vetor de dados originais X de um sistema multivariado de p variáveis em que $k < p$ componentes principais são usadas, pode ser expresso por

$$X = \bar{X} + VY + (X - \hat{X}), \quad (2.1.43)$$

em que o primeiro termo do lado direito da equação representa a média amostral, o segundo a contribuição das componentes principais (V é uma matriz $p \times k$ e Y é um vetor $k \times 1$ das componentes principais de X) e o termo final é a quantidade não explicada pelo modelo de componentes principais, o resíduo. Valores discrepantes (*outliers*) neste terceiro termo é uma indicação de que um vetor observação não pode ser caracterizado adequadamente pelo subconjunto de componentes principais que

foi escolhido. Isto ocorre porque poucas componentes principais foram retidas para produzir o modelo ou porque a observação é, realmente, um *outlier* do modelo, é também possível que, no caso do controle de qualidade, que a estrutura de covariância e seu espaço de vetores associados tenham mudado com o passar do tempo (JACKSON, 1991, p. 35).

O resíduo $(X - \hat{X})$ pode ser testado pela estatística da soma dos quadrados dos resíduos

$$Q = (X - \hat{X})'(X - \hat{X}), \quad (2.1.44)$$

que representa a soma dos quadrados da distância de $(X - \hat{X})$ ao espaço definido pelo modelo de k componentes principais. Q também pode ser escrita como

$$Q = \sum_{i=k+1}^p Z_i^2, \quad (2.1.45)$$

a soma ponderada dos quadrados das últimas $p - k$ componentes principais, com Z_i definido em (2.1.46).

Uma desvantagem no uso de ACP como método de monitoramento de um processo, apontada por Mastrangelo, Runger e Montgomery (1996), é que deslocamentos na média do processo, para um vetor que seja ortogonal as primeiras k componentes principais da matriz de covariância Σ , não são perceptíveis ao esquema de controle. Entretanto, isto pode não ter muita importância, dependendo do tipo de causa atribuível antecipada. Isto sugere a análise dos resíduos das componentes não retidas.

Para Mason e Young (2002), uma desvantagem do uso de ACP é que ela pode inibir um procedimento de interpretação direto (ou fácil) das componentes principais em termos das variáveis do processo uma vez que cada componente é uma combinação linear das variáveis do processo.

Jackson (1991, p. 55) apresenta vários critérios para determinação do número 'ótimo' de componentes principais a serem tomadas, que são os seguintes: Teste de Bartlett, proporção do traço explicado, variância residual individual, teste *scree*, teste da 'vara quebrada', raiz média (Guttman-Kaiser), método de Velicer, teste de validação cruzada; enquanto alguns tendem a superestimar, outros subestimam a

quantidade “correta”. A escolha do método para determinação pode depender das características dos dados disponíveis e dos objetivos específicos que se tem.

Uma ampla variedade de campos de aplicação da análise de componentes principais tem sido destacada em Jackson (1991, p. 58), dentre elas: psicologia, educação, controle de qualidade, química, pesquisa de mercado, economia, anatomia, biologia, silvicultura, agricultura.

Dado um vetor de variáveis aleatórias $\mathbf{X}' = (X_1, X_2, \dots, X_p)$ com matriz de covariância conhecida Σ e autovalores (por ser simétrica positiva definida) $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ e um vetor de coeficientes $\mathbf{a}'_i = (a_{i1}, a_{i2}, \dots, a_{ip})$, ao se considerar combinações lineares do tipo

$$\begin{aligned} Z_1 &= \mathbf{a}'_1 \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\ Z_2 &= \mathbf{a}'_2 \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\ &\vdots \\ Z_p &= \mathbf{a}'_p \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p. \end{aligned} \quad (2.1.46)$$

Usando propriedades da variância se mostra que

$$\begin{aligned} V(Z_i) &= \mathbf{a}'_i \Sigma \mathbf{a}_i, \\ \text{Cov}(Z_i, Z_k) &= \mathbf{a}'_i \Sigma \mathbf{a}_k. \end{aligned} \quad (2.1.47)$$

Para $i, k = 1, \dots, p$, se deseja construir as novas variáveis (componentes principais) Z_i de modo que tenham variância máxima possível e sejam não correlacionadas, restringindo que os vetores \mathbf{a}_i tenham norma unitária.

Segundo Montgomery (2013, p. 340), a interpretação geométrica das componentes principais Z_1, Z_2, \dots, Z_p é que estas correspondem aos eixos de um novo sistema de coordenadas no qual estas direções apresentam variabilidade máxima, em outras palavras, a componente principal Z_1 responde pela maior parte da variabilidade nas variáveis originais e assim por diante.

Na FIGURA 2.15, para o caso mais simples de duas variáveis, um conjunto de pontos consegue ter sua variabilidade máxima melhor visualizada no sistema de coordenadas (z_1, z_2) ao invés do sistema original (x_1, x_2) .

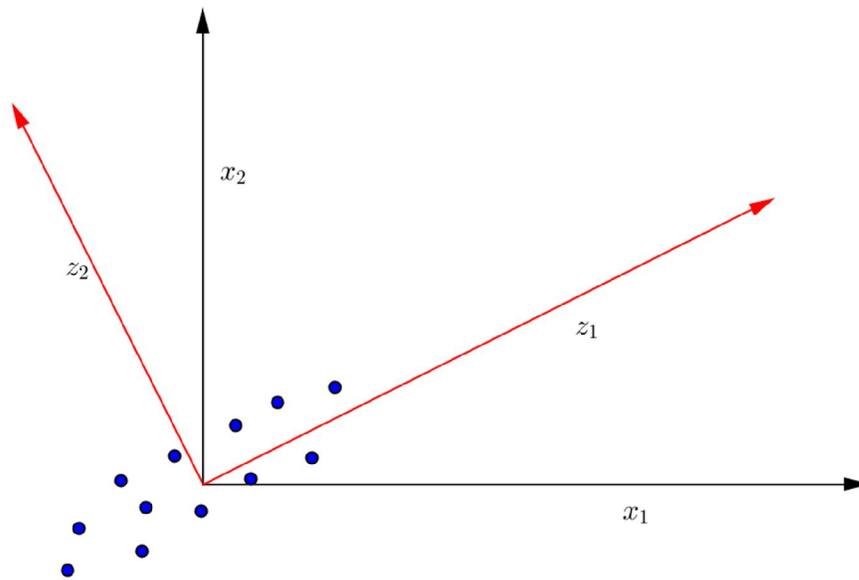


FIGURA 2.15 - Interpretação geométrica de componentes principais para $p=2$
 FONTE: O autor

Desta forma, se conduz a um problema de otimização em etapas de modo que a primeira componente principal consiste em determinar \mathbf{a}_1 que maximiza $V(Z_1) = \mathbf{a}'_1 \Sigma \mathbf{a}_1$ sujeito a $\mathbf{a}'_1 \mathbf{a}_1 = 1$. Em seguida, determinar \mathbf{a}_2 que maximiza $V(Z_2) = \mathbf{a}'_2 \Sigma \mathbf{a}_2$ sujeito a $\mathbf{a}'_2 \mathbf{a}_2 = 1$ e $Cov(Z_1, Z_2) = 0$ e assim sucessivamente. Johnson e Wichern (2007, p. 432) demonstram que, nestas condições, a i -ésima componente principal $i = 1, \dots, p$, é dada por

$$Z_i = \mathbf{e}_i' X = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p \quad (2.1.48)$$

e

$$V(Z_i) = \mathbf{e}'_i \Sigma \mathbf{e}_i = \lambda_i, \quad (2.1.49)$$

onde $(\lambda_i, \mathbf{e}_i)$ corresponde ao par de autovalor (em ordem decrescente de valor) e autovetor da matriz de covariância. Além disso,

$$\sum_{i=1}^p \sigma_{ii} = \sum_{i=1}^p V(X_i) = \lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_{i=1}^p V(Z_i). \quad (2.1.50)$$

Disso resulta que a proporção total de variância explicada pela k -ésima componente principal é

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad (2.1.51)$$

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

Uma vantagem das componentes principais é que dependem unicamente da matriz de covariância Σ e não exige a hipótese de normalidade. Por outro lado, dado um vetor de variáveis aleatórias $\mathbf{X}' = (X_1, X_2, \dots, X_p) \sim N_p(\boldsymbol{\mu}, \Sigma)$, as componentes principais $Z_i = \mathbf{e}_i' \mathbf{X}$ encontram-se nas direções de um elipsoide de densidade constante centrado na média, pois a partir da conhecida equação

$$(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = c^2, \quad (2.1.52)$$

pode-se chegar à equação abaixo que define um elipsoide com eixos nas componentes principais

$$\sum_{i=1}^p \frac{1}{\lambda_i} Z_i^2 = c^2. \quad (2.1.53)$$

De acordo com Johnson e Wichern (2007, p. 448), a partir de uma amostra de vetores \mathbf{X} , pode-se aproximar $\boldsymbol{\mu}$ por $\bar{\mathbf{X}}$ e Σ por S , respectivamente a média e variância amostral, os vetores \mathbf{X} de ordem $p \times 1$ que satisfazem a equação

$$(\mathbf{X} - \bar{\mathbf{X}})' S^{-1} (\mathbf{X} - \bar{\mathbf{X}}) = c^2, \quad (2.1.54)$$

constituem o contorno (curva de nível) de densidade constante $(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = c^2$ da distribuição normal subjacente.

Quando se padronizam as variáveis e trabalha-se com a matriz de correlação ao invés da matriz de covariância, os respectivos pares de autovalores e autovetores, em geral, não são os mesmos, nem mesmo existe uma função simples que os relaciona e é sugerido que as variáveis que estão em escalas de domínio e variação muito diferentes sejam padronizadas (JOHNSON e WICHERN, 2007, p. 449).

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 (MONTGOMERY, 2013).

Uma questão que não possui uma resposta definitiva é a respeito da quantidade de componentes principais a se considerar. Isso vai depender do caso em estudo. A escolha normalmente depende da quantidade de variância explicada, algum critério baseado na magnitude dos autovalores, como o critério de Kaiser que indica que se deve reter as componentes principais associadas com autovalores maiores do que 1, ou um critério gráfico como o denominado *scree plot*, que faz um gráfico dos pares de índice versus autovalores (em ordem decrescente), ou seja, $(1, \lambda_1), (2, \lambda_2), \dots, (p, \lambda_p)$ e a determinação do número de componentes é feita no ponto em que os autovalores são 'relativamente pequenos' e tem praticamente o mesmo valor. A FIGURA 2.16 exibe um exemplo de *scree plot*.

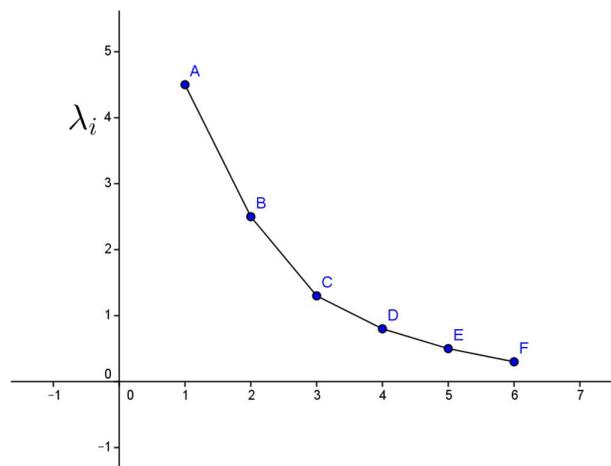


FIGURA 2.16 - Um típico *scree plot*
FONTE: O autor

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 componentes principais 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 (JACKSON, 1985, p. 2680).

A seleção do número de componentes principais a ser retida é subjetiva. Diversos critérios para a escolha também podem ser vistos em Jolliffe (2002) e Mastrangelo, Runger e Montgomery (1996) e Foucart (2000).

Um conjunto de n dados do espaço p dimensional (mesmo não provenientes da distribuição normal multivariada) pode ser expresso no sistema de coordenadas que coincide com os eixos do contorno de $(X - \bar{X})'S^{-1}(X - \bar{X}) = c^2$ que define um hiperelipsoide centrado em \bar{X} cujos eixos são os autovetores de S (ou, equivalentemente, de S^{-1}). Em outras palavras, as componentes principais podem ser vistas com o resultado de uma translação da origem do sistema de coordenadas para a média amostral \bar{X} , seguida de uma rotação dos eixos do sistema de coordenadas para que se sobreponham às direções de máxima variância. A FIGURA 2.17 exibe a ideia geométrica das componentes principais de um caso bivariado em que $\lambda_1 > \lambda_2$.

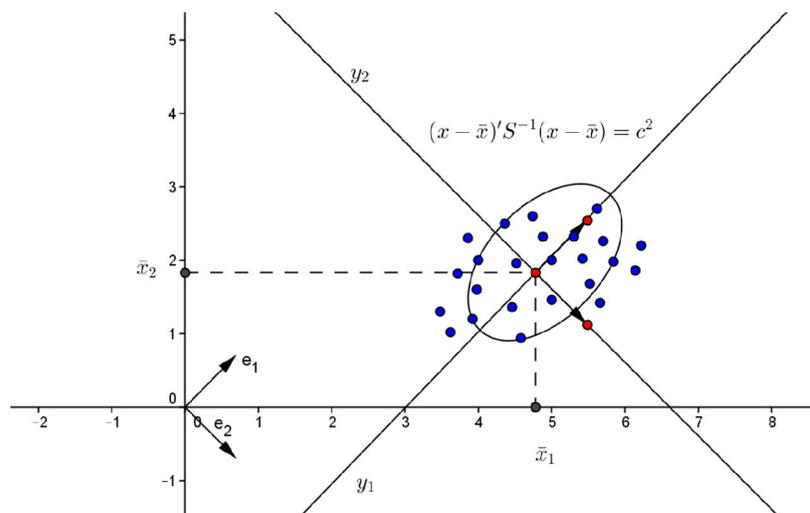


FIGURA 2.17 - Componentes principais e uma elipse com distância constante
FONTE: O autor

Para Johnson e Wichern (2007, p. 459) em situações onde há necessidade de monitorar a qualidade de um grande número de variáveis, o uso de componentes principais combinado com gráficos de controle multivariado é uma alternativa viável. Em um processo (envolvendo muitas variáveis) que se encontra estável no tempo no qual as características avaliadas são influenciadas apenas por causas comuns de variação, ou seja, que está sob controle estatístico, 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.

De uma amostra aleatória $X' = (X_1, X_2, \dots, X_p) \sim N_p(\mu, \Sigma)$, selecionadas k componentes principais $Z_k = e_k' X$ de uma amostra de n grande (≥ 50), Johnson e Wichern (2007, p. 460 e 464) propõe a construção do elipsoide

$$\sum_{i=1}^k \frac{Z_i^2}{\lambda_i} \leq \chi_{k, (1-\alpha)}^2, \quad (2.1.55)$$

onde os λ_i são autovalores de Σ e $\chi_{k, (1-\alpha)}^2$ corresponde ao valor crítico da distribuição χ^2 com k graus de liberdade para uma probabilidade $1 - \alpha$ que corresponderia a um gráfico baseado na estatística

$$T_k^2 = \sum_{i=1}^k \frac{Z_i^2}{\lambda_i}. \quad (2.1.56)$$

Relata também que é comum utilizar a estatística baseada na variação não explicada das variáveis originais, que tem distribuição χ_{p-k}^2 , por esta ser altamente efetiva em captar causas especiais de variação

$$T_{p-k}^2 = \sum_{i=k+1}^p \frac{Z_i^2}{\lambda_i}. \quad (2.1.57)$$

A FIGURA 2.18 mostra um exemplo simples em que duas variáveis fontes de causas comuns descrevem a variação de três variáveis do processo. Isto implica que as dimensões do modelo de componentes principais e do subespaço residual são dois e um, respectivamente.

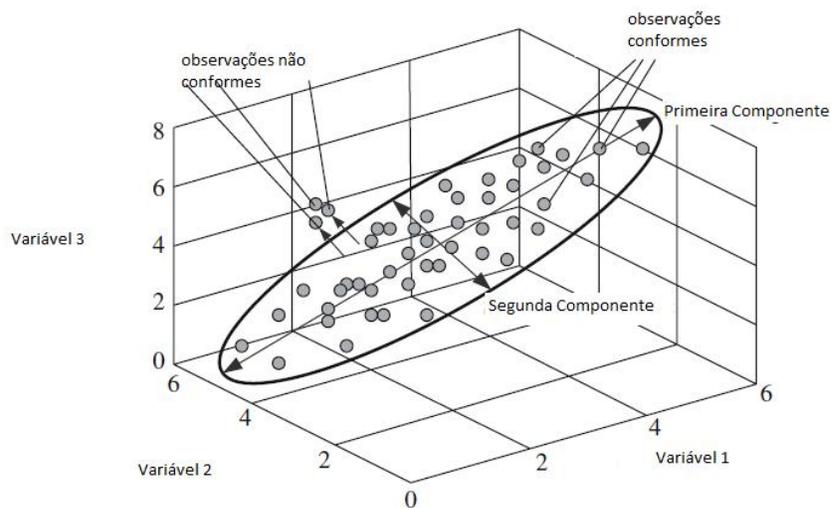


FIGURA 2.18 - Diagrama de um modelo de componentes principais e do seu subespaço complementar residual

FONTE: KRUGER e XIE (2012, p. 58)

2.1.5 Séries Temporais e Modelagem ARIMAX-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: economia, marketing, demografia, controle de processos (CHATFIELD, 1996). O objetivo é ajustar um modelo adequado aos dados disponíveis e capaz de ter um desempenho o mais preciso e acurado tanto quanto possível na previsão de valores da série.

Um modelo é uma representação simplificada da realidade que preserva uma equivalência adequada, buscando ser uma representação substitutiva da realidade. Um modelo se torna desejável quanto maior for seu poder de representatividade e o processo de verificação da representatividade de um modelo é a etapa de validação. Na validação se mede a adequabilidade do modelo à realidade representada (GOLDBARG e LUNA, 2000).

Uma série temporal (ST) é considerada uma realização de um *processo estocástico* que pode ser descrito como um fenômeno estatístico que se desenvolve no tempo de acordo com leis probabilísticas e que é caracterizado por uma equação estocástica (modelo), a partir da qual pode se obter alguns dos seus momentos (média e variância). Em processos estocásticos (P.E.) o futuro é incerto. Os modelos de processos estocásticos tentam capturar a estrutura de dependência existente no passado da série para realizar previsões. O grau de acerto das previsões depende assim de dois fatores chaves: i) que o modelo seja uma boa aproximação da estrutura de dependência da série (testes de especificação) e ii) que a estrutura de dependência identificada no passado permaneça razoavelmente estável no futuro (testes de previsão fora da amostra) (FERNANDES, 2006).

Os objetivos das séries temporais são: investigar seu mecanismo gerador, fazer previsões de valores futuros, descrever o seu comportamento, procurar periodicidades relevantes. Modelos probabilísticos ou estocásticos de séries temporais são construídos no domínio do tempo ou de frequências e devem ser o mais simples e parcimonioso possível (MORETTIN e TOLOI, 2006).

De acordo com Chatfield (1996), dentre os objetivos da análise de séries temporais estão a *descrição* através de um gráfico para obter medidas descritivas e

identificar padrões tais como: sazonalidade, tendências, possíveis *outliers*; a *explicação*, quando se usa uma série temporal para explicar outra (modelos de regressão múltipla); a *previsão* de valores futuros e o *controle* estatístico de qualidade de um processo.

A maioria dos métodos estatísticos trata de amostras aleatórias de observações independentes. Uma característica especial da análise de séries temporais é que observações sucessivas geralmente não são independentes e guardam uma autocorrelação, nesse caso os valores futuros podem ser previstos das observações passadas. Se a série pode ser prevista exatamente ela é dita ser determinística, mas na maioria das vezes apenas parcialmente se pode prever valores futuros dos valores passados e nesse caso ela é estocástica e se substitui a ideia da previsão exata dos valores futuros pela ideia que estes têm uma distribuição de probabilidades condicionada pelo conhecimento dos valores passados (CHATFIELD, 1996).

Diz-se que uma série temporal é estacionária se ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, reflexo de alguma forma de equilíbrio estável. As séries podem apresentar várias formas de não estacionariedade, como tendência linear, em que flutua ao redor de uma reta com inclinação não nula ou tendência exponencial.

Mais especificamente, costuma-se adotar estacionariedade de 2ª ordem (fraca) de uma série y_t se a média e a variância (incondicionais) do processo permanecem invariantes ao longo do tempo.

$$\begin{aligned} E(y_t) &= \mu, \forall t, \\ E[(y_t - \mu)^2] &= \gamma_0 = \sigma_y^2, \forall t, \\ E[(y_t - \mu)(y_{t-k} - \mu)] &= \gamma_k, \forall t. \end{aligned} \tag{2.1.58}$$

Alguns tipos de processos estocásticos úteis para o ajustamento de modelos para séries temporais são o processo puramente aleatório (ou ruído branco), o passeio aleatório, o processo de médias móveis, o processo autorregressivo, e os processos mistos e integrados de médias móveis e autorregressivos (CHATFIELD, 1996).

Segundo Fernandes (2006), a função de autocorrelação (FAC) representa uma medida de associação linear entre o presente do processo e o seu passado, ou seja,

uma medida da “memória” do processo. A FAC é o procedimento padrão para investigar a dependência linear relacionada por um modelo de ST. A ausência de autocorrelação de um PE não implica que este PE seja independente. Entretanto, se o PE for independente, a sua FAC, e qualquer outra medida de autoassociação será nula. A FAC de um processo estocástico é como se fosse uma “assinatura” deste processo. Processos de uma mesma “família” geram FAC’s do mesmo tipo. Portanto, a FAC pode ser utilizada como um procedimento para identificar o processo estocástico que gera uma série temporal. Se houver uma relação linear perfeita entre y_t e y_{t-k} , ou seja, entre uma observação no instante t e k instantes de tempo antes, então $|\rho_k| < 1$ e se $\rho_k = 0$, pode haver algum outro tipo de relação entre y_t e y_{t-k} (relação não-linear) onde

$$\rho_k = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}. \quad (2.1.59)$$

Em séries estacionárias o correlograma decai rapidamente para zero com o aumento de k ou é estatisticamente nulo para todo k . Em séries não estacionárias o correlograma decai lentamente para zero com o aumento de k . Outro instrumento para avaliar a estacionariedade são os testes de raízes unitárias, alguns destes testes estatísticos são: Dickey-Fuller (DF), *Augmented Dickey Fuller* (ADF) e Phillips-Perron (PP) (FERNANDES, 2006).

Alguns processos estocásticos não podem ser distinguidos através de suas FAC’s, pois estas apresentam formas similares. Este comportamento faz com que dada uma ST real, a partir da sua FAC estimada (correlograma) não se pode identificar se o modelo gerador da série é um AR(1) ou AR(2), por exemplo. É importante um tipo de função de autocorrelação que seja única para processos AR(p). Esta “função” existe e é denominada de FACP (Função de Autocorrelação Parcial). A dependência entre observações distantes de uma unidade de tempo faz com que, por transitividade, observações distantes k unidades de tempo sejam dependentes linearmente (FERNANDES, 2006).

Para estimar a “dependência líquida” entre observações distantes de k unidades de tempo, deve se eliminar da dependência total, a contribuição dada pelas observações intermediárias: $y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-k+1}$. O coeficiente de correlação

parcial calcula a correlação “líquida” entre y_t e y_{t-k} eliminando as influências de $y_{t-1}, y_{t-2}, y_{t-k+1}$. É possível mostrar que o coeficiente de correlação parcial de ordem k coincide com o último parâmetro de um modelo $AR(k)$. Por exemplo, para um processo $AR(2)$, $y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$, pode-se mostrar que o coeficiente de correlação parcial de ordem 2 é dado por:

$$\varphi_{22} = \text{corr}(y_t, y_{t-2} | y_{t-1}) = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}, \quad (2.1.60)$$

onde $\rho_k = \text{corr}(y_t, y_{t-k})$ e, mais geralmente, se conclui que a forma geral da FACP para um processo $AR(p)$ será $\varphi_k = \varphi_{kk} = 0, k > p$ e $\varphi_k = \varphi_{kk} \neq 0, k \leq p$. Esta é a FACP teórica, obtida a partir do modelo. A FACP de ordem k estimada a partir de uma série temporal é obtida estimando-se o último coeficiente de um processo $AR(k)$. Se o valor deste coeficiente estiver dentro do intervalo de confiança de 95%, então, ao nível de 5%, não podemos rejeitar a hipótese de que $\varphi_{kk} = 0$ (FERNANDES, 2006).

Modelos ARIMA (autorregressivos integrados de médias móveis) são capazes de descrever satisfatoriamente séries estacionárias e não estacionárias que não apresentem comportamento explosivo, caso em que, por exemplo, a série flutua ao redor de um nível, depois muda de nível e flutua ao redor de outro nível, e assim por diante (MORETTIN e TOLOI, 2006).

A maioria dos procedimentos em análise estatística de séries temporais presume a hipótese da estacionariedade. Quando esta não existir, uma alternativa é a transformação dos dados originais para obtenção de uma série estacionária. A transformação mais comum é tomar diferenças sucessivas da série original e denota-se a primeira diferença de $Z(t)$ por

$$\nabla Z(t) = Z(t) - Z(t - 1) \quad (2.1.61)$$

e a segunda diferença por

$$\nabla^2 Z(t) = \nabla(\nabla Z(t)) = \nabla(Z(t) - Z(t - 1)) = Z(t) - 2Z(t - 1) + Z(t - 2) \quad (2.1.62)$$

e, assim sucessivamente, normalmente é suficiente tomar uma ou duas diferenças para que a série se torne estacionária.

Todos os métodos de previsão são apenas diferentes procedimentos computacionais para calcular a previsão de mínimos quadrados de um valor futuro a partir de combinações lineares de valores passados e o modelo que descreve uma

série não leva necessariamente a um procedimento (ou fórmula) de previsão (MORETTIN e TOLOI, 2006). Portanto, é necessário especificar, além do modelo, uma função perda que é frequentemente o erro quadrático médio, em que, caso se possua observações de uma série temporal até o instante t e se deseje prever o valor da série no instante $t + h$, diz-se que $\hat{Z}_t(h)$ é a previsão de $Z(t + h)$, de origem t e horizonte h , define-se o erro quadrático médio de previsão como

$$E[Z(t + h) - \hat{Z}_t(h)]^2. \quad (2.1.63)$$

Uma fórmula para $\hat{Z}_t(h)$ é obtida quando se minimiza a equação acima.

Uma série temporal pode ser escrita na forma

$$Z(t) = f(t) + a_t, t = 1, \dots, N, \quad (2.1.64)$$

onde $f(t)$ é chamada sinal e a_t é o ruído.

Em modelos de erro ou de regressão, o sinal $f(t)$ é uma função que se pode determinar completamente, chamada de parte sistemática ou determinística, o ruído a_t é uma variável aleatória, independente de $f(t)$ em que os termos a_t são não correlacionados, tem média zero e variância constante, ou seja,

$$\begin{aligned} E(a_t) &= 0, \\ V(a_t) &= E(a_t^2) = \sigma_a^2, \\ E(a_t a_s) &= 0. \end{aligned} \quad (2.1.65)$$

Incluem-se aqui modelos de tendência linear como

$$Z_t = \alpha + \beta t + a_t \quad (2.1.66)$$

e outros modelos como curva de crescimento, polinomial, combinação de senos e cossenos em que os parâmetros podem ser estimados pelo método de mínimos quadrados. Em alguns casos a parte polinomial do sinal pode representar uma tendência e a parte trigonométrica uma sazonalidade.

A abordagem de Box e Jenkins (BOX, JENKINS e REINSEL, 2008) é uma metodologia bastante utilizada na análise de modelos paramétricos e consiste do ajuste de modelos autorregressivos integrados de médias móveis ($ARIMA(p, d, q)$) aos dados disponíveis. Sua construção se baseia num ciclo iterativo cujos estágios estão representados na FIGURA 2.19. No caso da análise dos resíduos indicar a estrutura do modelo como inadequada, deve-se retomar a etapa da identificação.

Muitas vezes são identificados vários modelos que passam pelos estágios de estimação e verificação e escolhe-se aquele que apresenta o menor erro quadrático médio de previsão.

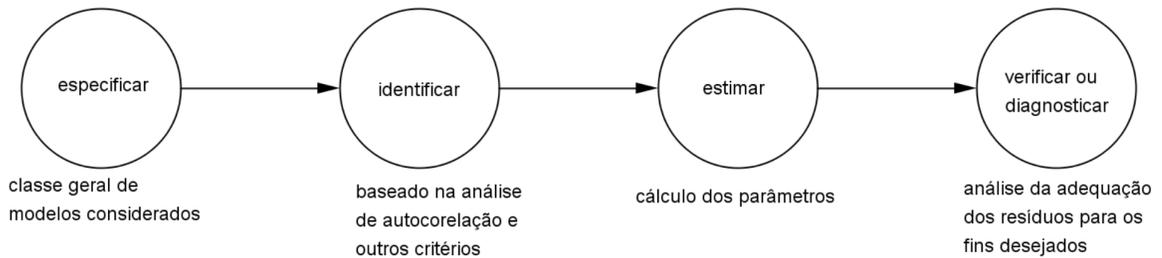


FIGURA 2.19 - Estágios do ciclo iterativo da construção de um modelo de séries temporais
FONTE: Adaptado de Morettin e Toloí (2006)

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 ∇ , de modo que $\nabla Z_t = Z_t - Z_{t-1} = (1 - B)Z_t$, os modelos lineares estacionários $ARMA(p, q)$ tem a forma

$$Z_t = \mu + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (2.1.67)$$

onde μ é a média, ϕ_i os coeficientes da parte autorregressiva e θ_i os coeficientes da parte de médias móveis. Com a notação do operador de retardo e , sem perda de generalidade, considerando a média 0, é costume se escrever

$$\phi(B)Z_t = \theta(B)a_t, \quad (2.1.68)$$

onde

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad (2.1.69)$$

e

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q. \quad (2.1.70)$$

Processos $AR(p)$, $MA(q)$ e $ARMA(p, q)$ apresentam características especiais em sua função de autocovariância (FAC) que são úteis no estágio da identificação de modelos adequados aos dados que se possui (MORETTIN e TOLOI, 2006). Estimativas das FAC que se acredita reproduzir as verdadeiras FAC desconhecidas são comparadas com tais características, descritas a seguir, para indicar modelos candidatos a melhor descrever a série temporal.

- $AR(p)$ tem FAC com decaimento exponencial e/ou de senoides amortecidas, infinita em extensão;
- $MA(q)$ tem FAC finita com corte na defasagem (*lag*) q ;
- $ARMA(p, q)$ tem FAC infinita que decai de acordo com exponenciais e/ou senoides amortecidas após a defasagem (*lag*) $q - p$.

Quando a série não se desenvolve ao longo do tempo ao redor de uma média constante, situação comum na prática, modelos ARMA não são apropriados para descrevê-las. Séries não estacionárias Z_t em que ao tomar-se um número finito de diferenças d estas se tornam estacionárias são chamadas de não estacionárias homogêneas. Seja

$$W_t = Z_t - Z_{t-1} = (1 - B)Z_t = \nabla Z_t \quad (2.1.71)$$

e, mais geralmente, se $W_t = \nabla^d Z_t$ é estacionária, então W_t pode ser representada por um modelo $ARMA(p, q)$, isto é,

$$\phi(B)W_t = \theta(B)a_t. \quad (2.1.72)$$

Quando W_t é uma diferença de Z_t , diz-se que Z_t é uma integral de W_t , por isso se diz que Z_t segue um modelo autorregressivo *integrado* de médias móveis ($ARIMA$) de ordem (p, d, q) ,

$$\phi(B)\nabla^d Z_t = \theta(B)a_t, \quad (2.1.73)$$

onde p e q são as ordem de autorregressão e de médias móveis, respectivamente.

Também é usual representar a equação (2.1.73) na forma

$$\varphi(B)Z_t = \underbrace{\phi(B)\nabla^d}_{\varphi} Z_t = \phi(B)(1 - B)^d Z_t = \theta(B)a_t. \quad (2.1.74)$$

Outro procedimento que facilita a identificação de séries temporais é a função de autocorrelação parcial (FACP) ϕ_{kk} que mede a correlação remanescente entre Z_t e Z_{t-k} após a eliminação da influencia de $Z_{t-1}, \dots, Z_{t-k+1}$. É conhecido que (BOX, JENKINS e REINSEL, 2008):

- Um modelo $AR(p)$ tem FACP $\phi_{kk} \neq 0$ para $k \leq p$ e $\phi_{kk} = 0$ para $k > p$;
- Um modelo $MA(q)$ tem FACP similar à FAC de um modelo $AR(p)$;
- Um modelo $ARMA(p, q)$ tem FACP que se comporta como a FAC de um MA puro.

A identificação de modelo(s) ARIMA para ajustar um conjunto de dados tem o objetivo de determinar os valores de p , d e q e estimar preliminarmente os parâmetros a serem usados na etapa de estimativas. A identificação é feita, sobretudo, com base nas autocorrelações e autocorrelações parciais estimadas, as quais se espera representarem adequadamente seus valores teóricos que são desconhecidos (MORETTIN e TOLOI, 2006).

As etapas da identificação são:

- Analisar a necessidade de aplicar transformação nos dados da série original com vistas a estabilizar sua variância, tornar os dados mais simétricos e próximos de uma distribuição normal;
- 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.

Foram criadas a partir da década de 70 formas alternativas de identificação de modelos $ARMA$, baseados numa função penalizadora de duas variáveis, que depende da estimativa da variância residual e de um termo penalizador que aumenta quando o número de parâmetros aumenta, de modo que o objetivo de minimizar consiste em encontrar o equilíbrio entre reduzir a variância residual e o número de parâmetros.

Alguns dos critérios de informação que são procedimentos de identificação que minimizam funções penalizadoras particulares, cujas fórmulas não serão apresentadas aqui, são: AIC – Critério de informação de Akaike, BIC – Critério de Informação Bayesiano, HQC – Critério de Hannan e Quinn, FPE – Critério '*final predictor error*', CAT – '*Criterion autoregressive transfer function*', dentre outros.

Após o ajuste de um modelo a uma série temporal é recomendável verificar se o modelo realmente fornece uma descrição adequada dos dados. Isto é feito através da análise do resíduo que é a diferença entre o valor da observação e o valor ajustado pelo modelo. Para um 'bom' modelo se esperam resíduos 'aleatórios' e próximos de zero em que cada coeficiente de autocorrelação é aproximadamente normalmente distribuído (CHATFIELD, 1996).

A aplicação de mínimos quadrados ordinários para estimar os coeficientes em uma equação cujos choques são autocorrelacionados, sem corrigir esta questão da autocorrelação, causa várias consequências infelizes: (1) O estimador de cada parâmetro não é melhor (no sentido de não ter variância mínima); (2) a variância residual provavelmente subestima a verdadeira variância; (3) o grau de ajuste R^2 é provavelmente superestimado; (4) as estatísticas usuais t e F são inválidas; (5) e previsões a partir da equação serão menos precisas do que poderiam ser desde que o padrão de autocorrelação nos resíduos contém informações úteis que estão sendo ignoradas (PANKRATZ, 1991).

Segundo Fernandes (2006), dentre os testes de diagnóstico dos resíduos mais utilizados na prática estão:

i) Teste de Jarque-Bera: A hipótese de normalidade dos resíduos num modelo ARIMA é importante, pois é crucial no processo de estimação pontual e das variâncias dos estimadores, intervalos de confiança e testes de hipóteses, são utilizadas nos testes de significância da FAC, FACP e intervalos de confiança. A normalidade será investigada através de dois coeficientes associados a distribuições de probabilidade, a assimetria – S (*Skewness*) e a Curtose – K (*kurtosis*). O teste de Jarque-Bera tem o objetivo de testar se uma dada ST provem de uma distribuição normal. É construído a partir da comparação das estimativas amostrais de S e K com os seus valores de referência sob normalidade, 0 e 3, respectivamente. Ela testa a hipótese $H_0: S = 0$ e $K = 3$ se os dados são normais e H_a : caso contrário. O teste é dado por:

$$JB = \frac{n\hat{S}^2}{6} + \frac{n(\hat{K} - 3)^2}{24} \sim \chi^2_{(2)}. \quad (2.1.75)$$

ii) Teste conjunto de autocorrelações Q de Ljung-Box (LB): Tem o objetivo de testar se as primeiras m autocorrelações de uma ST são nulas, ou seja, se a ST é não correlacionada. A hipótese do teste é $H_0: \rho_1 = \rho_2 = \dots = \rho_m = 0$ e H_a : pelo menos um dos $\rho_i \neq 0$ e a estatística do teste é

$$Q = T(T + 2) \sum_{k=1}^m \frac{\rho_k^2}{T - k} \sim \chi^2_{(v)}, \quad (2.1.76)$$

em que $v = m - (p + q + 1)$ são os graus de liberdade da distribuição χ^2 , ρ_k é a autocorrelação de *lag* k , T é a quantidade de dados e p, q são as ordens do modelo $ARMA(p, q)$.

iii) Teste de Breusch-Godfrey (BG) para autocorrelação: é uma alternativa ao teste LB. As vantagens do teste de BG sobre o teste LB em estudos de simulação mostram que o teste BG tende a ter mais sensibilidade em detectar a presença de autocorrelação nos dados e pode fornecer indicações sobre como o modelo deveria ser corrigido.

iv) Teste BDS (Brock, Dechert, Sheinkman) para a independência: O teste LB apenas testa se a série/os resíduos são não correlacionados. Aceitar a hipótese nula não necessariamente implica que a série dos resíduos esteja desprovida de qualquer tipo de dependência. O teste de BDS é um teste mais geral, onde a aceitação da sua hipótese nula implica na ausência de qualquer estrutura de dependência nos dados. No teste BDS tem-se que H_0 : os dados são *i.i.d.*; H_a : os dados apresentam dependência linear, não linear (média e/ou variância).

v) Teste do efeito GARCH: Este teste investiga se há presença de volatilidade, ou seja, mudanças na variância condicional da série, testando-a nos resíduos. Se houver evidência de volatilidade, então deve-se acoplar ao modelo ARIMA um modelo ARCH/GARCH para capturar as mudanças na variância condicional da série. É testada a hipótese H_0 : não existe efeito ARCH até ordem q nos resíduos, contra H_a : existe efeito ARCH.

Vi) Teste de Durbin-Watson: teste a hipótese nula de inexistência de autocorrelação contra a hipótese alternativa de autocorrelação residual de primeira ordem. É definido por

$$DW = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2}. \quad (2.1.77)$$

É possível mostrar que $DW \approx 2(1 - \rho_1)$, então, se não existe autocorrelação nos resíduos de ordem 1, $\rho_1 = 0$ e $DW \approx 2$ e podem ser obtidos limites inferior e superior para os valores críticos da distribuição do teste de DW (VOGELVANG, 2005).

A modelagem ARIMA tem destaque devido a ampla divulgação e utilização e superioridade em várias situações. Entretanto, muitas séries econômicas e financeiras apresentam heterocedasticidade condicional, isto é, a variância condicional varia com

tempo e, estas séries requerem modelos específicos para descrever a evolução da volatilidade no tempo, sugerem-se então os modelos ARCH (*Autoregressive Conditional Heteroscedasticity*) (MORETTIN e TOLOI, 2006).

Quando se tem séries com a variância evoluindo no tempo (condicional), modelos lineares como ARIMA não são adequados para descrever tal tipo de comportamento. Modelos não lineares são mais adequados neste caso (MORETTIN e TOLOI, 2006, p. 37). Por exemplo, modelos ARCH são não lineares no que se refere à variância. A variância condicional também é chamada de volatilidade.

Sejam $\mu_t = E(X_t|\mathcal{F}_{t-1})$ e $h_t = V(X_t|\mathcal{F}_{t-1})$ a média e a variância condicional de X_t onde \mathcal{F}_{t-1} é a informação até o instante $t - 1$, ou seja, $\{X_1, X_2, \dots, X_{t-1}\}$. Supondo que $\mu_t = 0$, segue que $h_t = V(X_t|\mathcal{F}_{t-1}) = E(X_t^2|\mathcal{F}_{t-1}) - [E(X_t|\mathcal{F}_{t-1})]^2 = E(X_t^2|\mathcal{F}_{t-1})$.

Modelos não lineares têm a forma

$$X_t = \underbrace{g(a_{t-1}, a_{t-2}, \dots)}_{\text{média}} + \overset{\text{choque}}{\tilde{a}_t} \underbrace{h(a_{t-1}, a_{t-2}, \dots)}_{\text{variância}}, \quad (2.1.78)$$

em que $g(\cdot)$ representa a média condicional e $h^2(\cdot)$ a variância condicional e a_t são as inovações ou choques aleatórios supostamente i.i.d.. Se g for não linear o modelo se diz não linear na média e se h^2 for não linear o modelo se diz não linear na variância. Por exemplo, o modelo

$$X_t = \alpha a_{t-1}^2 + a_t, \quad (2.1.79)$$

em que $\alpha \geq 0$ é uma constante, $g(a_{t-1}, a_{t-2}, \dots) = \alpha a_{t-1}^2$ e $h(a_{t-1}, a_{t-2}, \dots) = 1$ é não linear na média e linear na variância, enquanto que o modelo *ARCH(1)* dado por

$$X_t = a_t \sqrt{\alpha X_{t-1}^2}, \quad (2.1.80)$$

em que $g(a_{t-1}, a_{t-2}, \dots) = 0$ e $h(a_{t-1}, a_{t-2}, \dots) = \sqrt{\alpha X_{t-1}^2}$, $h^2(a_{t-1}, a_{t-2}, \dots) = \alpha X_{t-1}^2$, dado que $X_{t-1} = f(a_{t-1})$, é linear na média e não linear na variância. Note que $h^2(\cdot) = \alpha X_{t-1}^2$ é a variância condicional (não constante).

Os modelos autorregressivos com heterocedasticidade condicional (ARCH) foram introduzidos por Engle (1982). Sua ideia básica é que não exista autocorrelação serial, mas que exista variância condicional (volatilidade) por meio de uma função quadrática dos dados passados (MORETTIN e TOLOI, 2006, p. 363).

A forma geral de um modelo $ARCH(r)$ é

$$X_t = \varepsilon_t \sqrt{h_t} = \varepsilon_t \tilde{h}_t, \quad (2.1.81)$$

onde

$$\begin{aligned} h_t &= \alpha_0 + \alpha_1 X_{t-1}^2 + \dots + \alpha_r X_{t-r}^2, \\ \tilde{h}_t &= \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2 + \dots + \alpha_r X_{t-r}^2} \end{aligned} \quad (2.1.82)$$

e ε_t é uma seqüência de variáveis aleatórias i.i.d. com média 0 e variância 1, $\alpha_0 > 0$ e $\alpha_i \geq 0$ para $i > 0$. Em geral se supõe que $\varepsilon_t \sim N(0,1)$ ou $\varepsilon_t \sim t_\nu$. Comparado a forma geral dos modelos não lineares indicada anteriormente, $g(\cdot) = 0$, ou seja, $ARCH(r)$ é linear na média e a considera 0 e $\tilde{h}_t^2 = h_t$ é não linear na variância.

Em particular num $ARCH(1)$, onde $X_t = \varepsilon_t \sqrt{h_t}$, $h_t = \alpha_0 + \alpha_1 X_{t-1}^2$, $\alpha_0 > 0$ e $\alpha_1 \geq 0$, tem-se que

$$\begin{aligned} E(X_t) &= E(X_t | \mathcal{F}_{t-1}) = 0, \\ V(X_t) &= V(X_t | \mathcal{F}_{t-1}) = E(X_t^2 | \mathcal{F}_{t-1}) = E(\alpha_0 + \alpha_1 X_{t-1}^2) = \alpha_0 + \alpha_1 E(X_{t-1}^2). \end{aligned} \quad (2.1.83)$$

Se o processo $\{X_t\}$ é estacionário de segunda ordem, então para todo t , $E(X_{t-1}^2) = E(X_t^2) = V(X_t)$, portanto

$$V(X_t) = \alpha_0 + \alpha_1 V(X_t) \rightarrow V(X_t) = \frac{\alpha_0}{1 - \alpha_1}. \quad (2.1.84)$$

Além disso, uma vez que $V(X_t) > 0$, $\alpha_0 > 0$, segue que $0 \leq \alpha_1 < 1$.

No modelo $GARCH(r,s)$ ($ARCH$ generalizado), proposto por Bollerslev (1986), o choque aleatório a_t , em (2.1.67), é representado por:

$$a_t = \sigma_t \varepsilon_t, \quad (2.1.85)$$

onde σ_t é sua variância, que agora não é mais constante, mas sim, possui volatilidade e ε_t é o termo estocástico. A volatilidade significa que a variância σ_t é autodependente dos choques quadráticos e variâncias anteriores, de modo que, segundo Gujarati e Porter (2010), a variância em (2.1.85) é dada por

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^r \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^s \beta_i \sigma_{t-i}^2, \quad (2.1.86)$$

onde r e s são as respectivas ordens de dependência dos choques quadráticos e variâncias anteriores. Finalmente, um modelo composto pelas equações (2.1.67), (2.1.85) e (2.1.86) é denominado de ARIMA-GARCH.

Para avaliar a estacionariedade da série temporal é usual fazer uma inspeção visual da série em nível e observar se ela é ou não é estacionária e para se certificar da estacionariedade sugere-se fazer um teste de raiz unitária (Dickey-Fuller).

Para a identificação, se estimam os coeficientes de autocorrelação e autocorrelação parcial e a estatística Q de Ljung-Box dos resíduos. Os coeficientes significativos da FAC sugerem termos MA e da FACP, termos AR . Para o diagnóstico de uma série temporal, dentre alguns modelos selecionados na identificação, deve-se escolher aquele que possuir o menor critério entre Akaike e Schwarz. Analogamente para os termos GARCH(r,s) deve-se analisar os coeficientes de autocorrelação e autocorrelação parcial e a estatística Q de Ljung-Box dos resíduos quadráticos.

Séries temporais modeladas com a metodologia $ARMA$ ($ARIMA$) de Box & Jenkins estão baseados somente no passado de sua própria variável para fins de previsões, ou seja, não se baseiam em nenhuma teoria econômica ou informação da equipe de engenharia, portanto, seus coeficientes não têm uma interpretação específica. O exame da plausibilidade do modelo é feito como um todo, verificando se ele descreve bem os dados e se produz boas previsões.

Seja $y_t (t = 1, \dots, T)$ uma série temporal estacionária (ou não estacionária que possa tornar-se estacionária mediante algum tipo de transformação como, por exemplo, diferenciações) que possui estrutura de autodependência linear. Se $\left((x_{1,t})_{t=1}^T, \dots, (x_{r,t})_{t=1}^T \right)$ representam um conjunto de vetores de realizações de r variáveis exógenas de $y_t (t = 1, \dots, T)$ então, segundo Box e Tiao (1975) e Pankratz (1991), cada realização y_t pode ser representada por um modelo ARIMAX (p, d, q), ARIMA com variáveis exógenas ou explicativas, com a seguinte formulação:

$$\nabla^d y_t = \phi_0 + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^q \theta_j a_{t-j} + \sum_{i=1}^r \sum_{k=0}^l \alpha_{ik} x_{i,t-k} + a_t, \quad (2.1.87)$$

onde B é o operador de retardo, definido por $B^m y_t := y_{t-m}$, $m \in \mathbb{Z}$; $\nabla^d := (1 - B)^d$ é o operador diferença de ordem d ; $(\phi_i)_{i=0}^p$ e $(\theta_j)_{j=1}^q$ são os conjuntos

de parâmetros autorregressivos e de medias móveis, respectivamente de ordem p e q , em que $\phi_p, \theta_q \neq 0$ e $(\alpha_{ik})_{k=0}^l$ é o conjunto de parâmetros associados com a i -ésima componente exógena (todos estes parâmetros necessitam satisfazer as condições de estacionariedade e invertibilidade (veja, por exemplo, Hamilton (1994) e (Lutkepohl (2006))); a_t é um choque (ou inovação) que consiste de uma realização de uma variável aleatória ε_t proveniente de um processo estocástico independente e identicamente distribuído $\varepsilon_t (t = 1, \dots, T)$, com distribuição padrão, isto é, média 0 e variância constante e igual a 1, geralmente, distribuição normal, t -Student ou GED – *generalized error distribution*. Em particular, se $\alpha_{ik} = 0$ para todo $i = 0, \dots, r$ e todo $k = 0, \dots, l$, então o modelo exibido em (2.1.86) torna-se o ARIMA (p, d, q) convencional, ou seja, esta formulação de um modelo de regressão linear, muitas vezes chamada de regressão linear dinâmica, é uma generalização dos modelos ARIMA. Por sua vez, um modelo SARIMAX $(p, d, q) \times (P, D, Q)_S$ é usado para modelar series temporais sazonais (HAMILTON, 1994) e consiste de um modelo ARIMAX com componentes sazonais.

A previsão de uma série temporal pode ser *ex-ante* (estática) quando é feita para calcular valores futuros de curto prazo da variável em estudo, ou *ex-post* (dinâmica) para gerar valores dentro do período amostral, quanto melhor for este último, mais eficiente o modelo estimado. Esta diferenciação ocorre quando se tem que escolher entre o valor previsto ou observado a ser usado no cálculo de previsões vários períodos à frente. O procedimento estático calcula previsões um período à frente (VOGELVANG, 2005).

Por exemplo, assumindo o seguinte modelo estimado

$$\hat{Y}_t = \alpha_1 \hat{Y}_{t-1} + \hat{\beta}. \quad (2.1.88)$$

As previsões podem ser calculadas de duas formas, primeiro se prevê Y_{n+1}

$$\hat{Y}_{n+1} = \alpha_1 Y_n + \hat{\beta}, \quad (2.1.89)$$

onde Y_n é a última observação de Y no período dentro da amostra de estimação.

Assim, Y_{n+2} pode ser previsto de duas formas

estática	$\hat{Y}_{n+2} = \alpha_1 Y_{n+1} + \hat{\beta},$	(2.1.90)
ou dinâmica	$\hat{Y}_{n+2} = \alpha_1 \hat{Y}_{n+1} + \hat{\beta}.$	

Usando Y_{n+1} (caso seja conhecido – período fora da amostra) ou usando \hat{Y}_{n+1} .

Para a análise da previsão deve-se ater a alguma medida da eficiência da previsão como o erro quadrático médio (EQM) ou *mean square error* (MSE), que é a média do quadrado da diferença entre cada valor previsto *ex-post* (previsão dinâmica) e o valor real observado na amostra. Ele é uma medida formal da qualidade das previsões *ex-post*. Quanto menor o MSE melhor será o grau de ajustamento do modelo aos dados da série temporal.

2.1.6 Decomposição *Wavelet* de uma Série Temporal

Quando uma função desconhecida para representar n pares de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ não é assumida ter qualquer forma paramétrica particular ou representação, mas sim, pertencer a uma classe de funções que possuem características mais gerais, tais como certo nível de ‘suavidade’, usando os dados observados, pode-se estimar tal função representando a função em outro domínio. Uma maneira comum de abordar isso é usar uma representação em série ortogonal da função. Isso muda o problema de estimação de diretamente tentar estimar a função desconhecida f , para estimar um conjunto de coeficientes escalares que representam f no domínio das séries ortogonais. Um método eficiente para estimar tais funções envolve a utilização de *Wavelets*, que são fortes ferramentas em tais métodos porque concentram a maioria das informações sobre a função em um conjunto muito reduzido de dados e têm a capacidade de estimar características globais e locais da função subjacente (HOLLANDER, WOLFE e CHICKEN, 2014, p. 629).

Uma *Wavelet* é uma função capaz de decompor e representar outra função (ou uma série de dados) originalmente descrita no domínio do tempo, de forma a permitir a análise desta função em diferentes escalas de frequência e de tempo. A decomposição de uma função com o uso de *Wavelets* é conhecida como *transformada Wavelet* e tem suas variantes contínuas e discretas. Graças à capacidade de decompor as funções tanto no domínio da frequência quanto no domínio do tempo, as funções *Wavelet* são ferramentas poderosas de processamento

de sinais, muito aplicadas na eliminação de ruído, separação de componentes do sinal, identificação de singularidades e previsão de séries temporais.

Uma *Wavelet* é uma função com um número determinado de oscilações que duram por certo período de tempo e tendem a zero no resto do período. Esta característica permite as *Wavelets* analisarem séries temporais não estacionárias, ao contrário das funções trigonométricas na série de Fourier.

Uma ‘família’ *Wavelet* tem uma *Wavelet* ‘mãe’ ψ que produz *Wavelets* ‘filhas’ $\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$ por meio de expansões (dilatações) e translações, u é o índice de localização que indica posição no domínio do tempo e s é a largura da *Wavelet* filha (KRIECHBAUMER *et al.*, 2014).

Para ser considerada uma *Wavelet*, uma função ψ tem que atender as seguintes características:

- A área total sob a curva da função é 0, ou seja $\int_{-\infty}^{\infty} \psi(t) dt = 0$ e
- A energia da função é finita, ou seja, $\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty$.

Estas condições são equivalentes a dizer que $\psi(t)$ é quadrado integrável ou que pertence ao conjunto $\mathcal{L}^2(\mathbb{R})$ das funções quadrado integráveis. As propriedades acima sugerem que $\psi(t)$ tende a oscilar acima e abaixo do eixo t , e que tem sua energia localizada em certa região, já que ela é finita.

Essa característica de energia concentrada em uma região finita é que diferencia a análise usando *Wavelets* da análise de Fourier, já que esta última usa as funções de seno e cosseno que são periódicas e infinitas. Outra forma de expressar esta característica é dizer que a transformada *Wavelet* é um operador local no domínio do tempo. A FIGURA 2.20 e a FIGURA 2.21 exibem esta característica citada.

Alguns exemplos de bases *Wavelets* discretas para o espaço $\mathcal{L}^2(\mathbb{R})$ são: Coiflets, Cohen-Daubechies-Feauveau, Daubechies, Haar, Mathieu, Legendre, Villasenor, Symlets e de *Wavelets* contínuas reais: Beta, Hermitian, Hermitian *hat*, Mexican *hat*, Meyer e Shannon e complexas: Complex Mexican *hat*, Morlet.

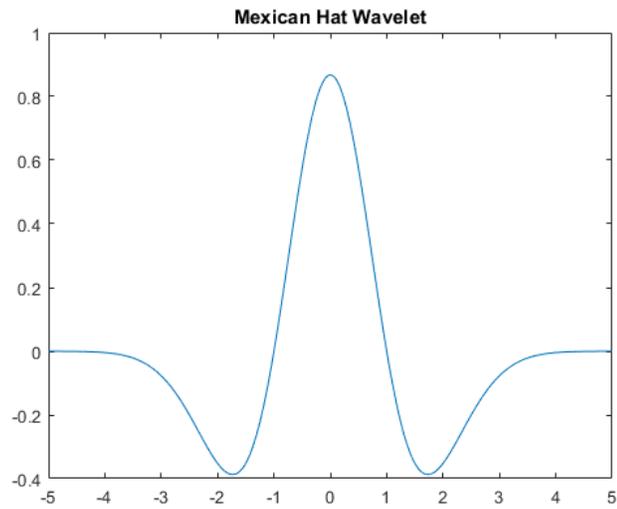


FIGURA 2.20 – Uma Wavelet tipo chapéu mexicano
FONTE: MATHWORKS (2016)

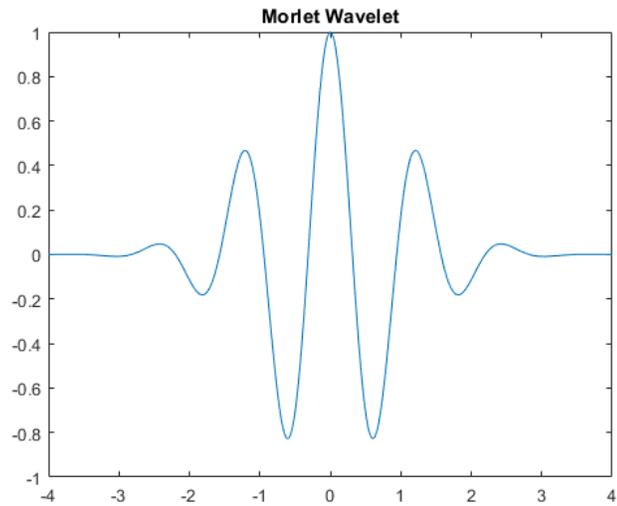


FIGURA 2.21 – Uma Wavelet de Morlet
FONTE: MATHWORKS (2016)

As transformadas *Wavelet* são usadas para decompor uma série temporal numa série (ou componente) de aproximação A_j , com seus coeficientes $a_{j,k}$ e um conjunto de séries (componentes) de detalhes D_j com seus coeficientes $d_{j,k}$. A componente de aproximação é a principal componente e pode ser considerada uma versão ‘filtrada’ (alisada, sem ruído) da série original. Séries de detalhes captam flutuações da série original ao redor da componente de aproximação. A série original

pode ser reconstruída através da soma da componente de aproximação e das componentes de detalhes (KRIECHBAUMER *et al.*, 2014).

Na FIGURA 2.22

FIGURA 2.22 - Decomposição Wavelet de nível 2 da série temporal de um instrumento de Itaipu, é apresentada a série temporal mensal de um instrumento de monitoramento em Itaipu, um pêndulo direto (PD-E-6) no período de 1992 a 2014 (280 observações), e os gráficos da decomposição *Wavelet* de nível 2 em que se tem a série de aproximação (A1) e duas séries de detalhes (D1 e D2) para a base de Haar11.

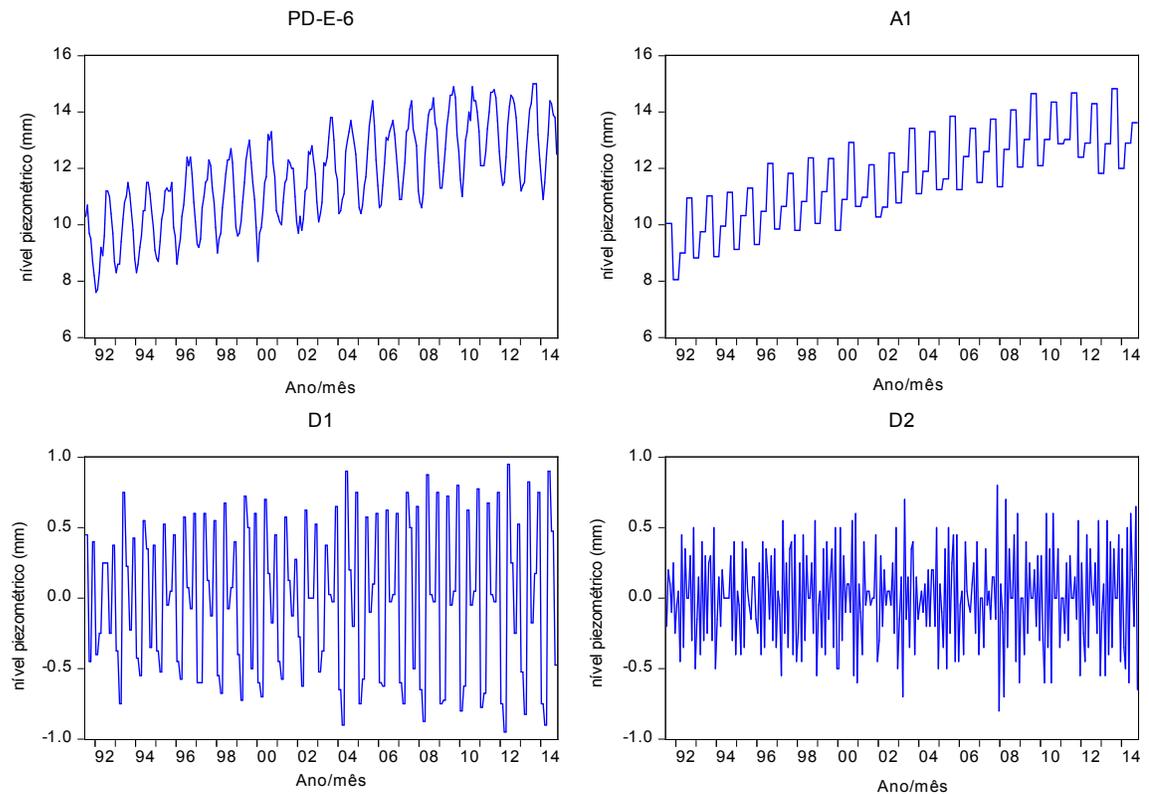


FIGURA 2.22 - Decomposição Wavelet de nível 2 da série temporal de um instrumento de Itaipu
FONTE: O autor

Algumas bases *Wavelet* bastante conhecidas são a de Haar e as de Daubechies. Segundo Hollander, Wolfe e Chicken (2014, p. 630), se ψ é uma função *Wavelet*, então o conjunto

$$\Psi = \{\psi_{jk} : j, k \in \mathbb{Z}\}, \quad (2.1.91)$$

onde

$$\psi_{jk} = 2^{j/2} \psi(2^j x - k), \quad (2.1.92)$$

é uma base para as funções quadrado integráveis e Ψ é um conjunto de translações e dilatações de ψ . A função ψ pode ser construída para garantir que o conjunto Ψ seja ortonormal. A propriedade $\int_a^b \psi^2 = 1$ implica que o valor de ψ está próximo de 0, exceto em pequeno domínio. Isto, combinado com a equação (2.1.92), significa que à medida que j aumenta, ψ_{jk} torna-se cada vez mais localizada. É possível ainda representar pelo somatório duplo

$$f(x) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x) \quad (2.1.93)$$

de *Wavelets* produzidas pelas translações e dilatações de ψ . Isto pode dar a interpretação de uma série de aproximações para f , onde cada aproximação está em um nível de escala diferente j , em que a escala pode ser visto como sinônimo de frequência. Nas *Wavelets* o termo resolução é usado para se referir à escala ou frequência, para um j fixo, o índice k representa o comportamento de f na escala de resolução e numa localização particular. A notação acima pode ser interpretada como dando informações sobre a função f em diferentes níveis de resolução (escala, frequência) j e localização k (HOLLANDER, WOLFE e CHICKEN, 2014, p. 631).

Representando

$$f_j(x) = \sum_{j < J} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x), \quad (2.1.94)$$

quando j aumenta, f_j permite modelar o comportamento de f em escala menor (frequência superior). Isto corresponde a mudanças em f que ocorrem sobre um pequeno intervalo do eixo x . Já quando j diminui, f_j modela em escala maior (frequência inferior) o comportamento de f .

Também é comum escrever

$$f(x) = \sum_{k \in \mathbb{Z}} \xi_{j_0 k} \phi_{j_0 k}(x) + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x), \quad (2.1.95)$$

em que o primeiro termo do lado direito da equação (2.1.95) é a aproximação acumulada f_{j_0} usando todos os níveis de resolução $j < j_0$ relacionada com a *Wavelet* ψ através de

$$f_{j_0}(x) = \sum_{j < j_0} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x) = \sum_{k \in \mathbb{Z}} \xi_{j_0 k} \phi_{j_0 k}(x), \quad (2.1.96)$$

onde

$$\begin{aligned} \phi_{j_0 k}(x) &= 2^{j_0/2} \psi(2^{j_0} x - k) \\ \xi_{j_0 k} &= \langle f, \phi_{j_0 k} \rangle. \end{aligned} \quad (2.1.97)$$

A aproximação f_{j_0} é considerada “suave” ou “grosseira” com relação a uma aproximação que incluiria níveis de resolução superiores $j \geq j_0$. Ainda, o primeiro termo da direita da equação (2.1.95) representa uma aproximação suave de f e a segunda série permite a modelagem do comportamento de f em uma escala de frequência superior (HOLLANDER, WOLFE e CHICKEN, 2014, p. 632).

A primeira, portanto, é chamada de série de aproximação e a segunda chamada de série de detalhes, que quando adicionada a primeira, cria uma menos suave e mais detalhada aproximação de f em níveis de resolução superiores.

Dado o par ordenado $(l^2, \langle ; \rangle)$, onde l^2 é o conjunto das seqüências infinitas quadrados somáveis de números complexos, isto é $l^2 := \{f: \mathbb{Z} \rightarrow \mathbb{C} / \sum_{t \in \mathbb{Z}} |f(t)|^2 < \infty\}$ e a função $\langle ; \rangle: l^2 \rightarrow \mathbb{C}$ que é um produto interno. Um elemento $\omega(\cdot)$ em l^2 é um vetor l^2 – *Wavelet* com produto interno $\langle ; \rangle: l^2 \rightarrow \mathbb{C}$, se e somente se a sequência duplamente indexada $\{\omega_{m,n}(\cdot)\}_{(m,n) \in \mathbb{Z} \times \mathbb{Z}}$ consiste de uma base ortonormal para l^2 , onde o parâmetro m é chamado de parâmetro de escala e n , é chamado de parâmetro de translação. Por sua vez, um elemento $\phi(\cdot) \in l^2$ é um vetor de escala para o produto interno $\langle ; \rangle: l^2 \rightarrow \mathbb{C}$, se, e somente se, o conjunto $\{\phi_{m,n}(\cdot)\}_{(m,n) \in \mathbb{Z} \times \mathbb{Z}}$, para todo $m, n \in \mathbb{Z}$, dado por $\phi_{m,n}(\cdot) = 2^{m/2} \phi(2^m(\cdot) - n)$ é tal que $\langle \phi_{l,i}(\cdot); \phi_{j,k}(\cdot) \rangle = 0$, para $l = j$ e $i \neq k$ e $\langle \phi_{l,i}(\cdot); \phi_{j,k}(\cdot) \rangle \neq 0$, caso contrário.

Segundo Kubrusly (2011), Mallat (1998) e Kubrusly e Levan (2006), dada uma função $f(\cdot) \in l^2$, $f(\cdot)$ pode ser decomposta ortogonalmente unicamente em termos de

uma base *Wavelet* ortonormal, esta base pode ser representada por $\{\phi_{m_0,n}(\cdot)\}_{n \in \mathbb{Z}} \cup \{\omega_{(m,n)}(\cdot)\}_{(m,n) \in \{m\}_{m=m_0}^{+\infty} \times \mathbb{Z}}$, como em

$$f(\cdot) = f_{V_{m_0}(\phi)}(\cdot) + \sum_{m=m_0}^{+\infty} f_{W_m(\omega)}(\cdot). \quad (2.1.98)$$

Onde: $f_{V_{m_0}(\phi)}(\cdot) := \sum_{n \in \mathbb{Z}} a_{m_0,n} \phi_{m_0,n}(\cdot)$ é a componente de aproximação de nível m_0 , $a_{m_0,n} := \langle f(\cdot), \phi_{(m_0,n)}(\cdot) \rangle$ (isto é, o produto interno usual entre o vetor sinal $f(\cdot)$ e o vetor escala de nível m_0 e n); e $f_{W_m(\omega)}(\cdot) := \sum_{n \in \mathbb{Z}} d_{m,n} \omega_{m,n}(\cdot)$ é a componente de detalhe de nível m , $d_{m,n} := \langle f(\cdot), \omega_{(m,n)}(\cdot) \rangle$ (isto é, o produto interno usual entre $f(\cdot)$ e o vetor *Wavelet* de nível m e n). Os coeficientes $a_{m_0,n}$ e $d_{m,n}$ são, respectivamente, os coeficientes *Wavelet* de aproximação e de detalhe. Na equação (2.1.98), existe uma decomposição *Wavelet* de $f(\cdot)$ sobre l^2 em termos da base $\{\phi_{m_0,n}(\cdot)\}_{n \in \mathbb{Z}} \cup \{\omega_{(m,n)}(\cdot)\}_{(m,n) \in \{m\}_{m=m_0}^{+\infty} \times \mathbb{Z}}$.

Seja $\chi_{[0,T]}$ a função tal que $\chi_{[0,T]} := 1$, se $t \in \{1, \dots, T\}$ e $\chi_{[0,T]} := 0$ se $t \in \mathbb{Z} - \{1, \dots, T\}$ e $y(\cdot) := (y_t)_{t=1}^T$ é uma série temporal de tamanho T , onde $T > 1$. Assim, a composição $\chi_{[0,T]} \circ y(\cdot) = \tilde{y}(\cdot): \mathbb{Z} \rightarrow \mathbb{R}$ pode ser vista como uma sequência $\tilde{y}(\cdot) := (\dots, 0, 0, y_1, y_2, \dots, y_T, 0, 0, \dots)$ em l^2 . Assumindo que $\{\phi_{m_0,n}(\cdot)\}_{n \in \mathbb{Z}} \cup \{\omega_{(m,n)}(\cdot)\}_{(m,n) \in \{m\}_{m=m_0}^{+\infty} \times \mathbb{Z}}$ é uma base ortonormal *Wavelet* para l^2 , segue que a série temporal $y(\cdot)$ admite ser decomposta, aproximadamente conforme dado em (2.1.99):

$$\tilde{y}(\cdot) \simeq \tilde{\tilde{y}} = \sum_{n=1}^{n_{m_0}} a_{m_0,n} \phi_{m_0,n}(\cdot) + \sum_{m=m_0}^{m_0+(p-1)} \sum_{n=1}^{n_m} d_{m,n} \omega_{m,n}(\cdot). \quad (2.1.99)$$

A expansão em (2.1.99) é chamada de decomposição *Wavelet* de nível p . O valor adotado para o parâmetro de nível m_0 é geralmente o mesmo de p , isto é, $m_0 = p$. Desta forma, dada uma série temporal $y_t, (t = 1, \dots, T)$, a decomposição ortogonal *Wavelet* de nível p de y_t consiste de

$$y_t \cong y_{A_p,t} + \sum_{m=p}^{p+(p-1)} y_{D_m,t}, \quad (2.1.100)$$

onde $p \in \mathbb{Z}$ é o parâmetro de nível de escala (resolução ou frequência), y_{A_p} é a componente *Wavelet* (CW) de aproximação de nível p , y_{D_m} é uma CW de detalhe de

nível $p, p + 1, \dots$, e $y_{A_p,t} = \sum_{n \in \mathbb{Z}} a_{p,n} \phi_{p,n}(t)$ e $y_{D_m,t} = \sum_{n \in \mathbb{Z}} d_{m,n} \omega_{m,n}(t)$, onde $a_{p,n}$ é o coeficiente de aproximação de nível p e n , ϕ é a função escala ('pai'), $d_{m,n}$ é o coeficiente de detalhe de nível m e n , ω é a função *Wavelet* ('mãe'). Para um m_0 fixo, os coeficientes de aproximação e de detalhes são dados pelo produto interno entre a série temporal y_t e as funções escala e *Wavelets* filhas.

A *Wavelet* de Haar $\psi = \psi_{00}$ é dada por

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2) \\ -1, & x \in [1/2, 1) \\ 0, & \text{caso contrário.} \end{cases} \quad (2.1.101)$$

A transformada discreta *Wavelet* produz somente o número mínimo de coeficientes necessário para reconstruir a série original, discretizando os parâmetros u, s , de modo que $u = k2^{-j}$ e $s = 2^{-j}$, onde $j, k \in \mathbb{Z}$ e $j = 1, \dots, J$ onde J é o nível de decomposição e $T = 2^J$ é o número de dados da série, ou seja, a transformada discreta se restringe a séries diádicas. Outros tipos de ajustes podem ser feitos para trabalhar com qualquer quantidade de dados. Uma componente de detalhes indica a escala de tempo das flutuações. Por exemplo, numa série mensal, o nível de decomposição D_1 indica uma escala de tempo de flutuação de 2 a 4 meses, o nível D_2 de 4 a 8 meses e assim por diante (KRIECHBAUMER *et al.*, 2014).

Detalhes matemáticos sobre a ortonormalidade da base de Haar e que existe um nível de decomposição em que se pode aproximar tanto quanto se queira uma função contínua por um conjunto de *Wavelets* de Haar podem ser encontrados em Walnut (2002).

2.1.7 Redes Neurais Artificiais

Em meados da década de 80 aumentou o interesse nas Redes Neurais Artificiais (RNA's) por causa do avanço tecnológico e das soluções obtidas em diversos tipos de problemas. Uma rede neural artificial é uma forma de computação caracterizada por sistemas que lembram a estrutura do cérebro humano. Possuem unidades de processamento simples (nós - neurônios) nos quais são calculadas determinadas funções matemáticas, geralmente não lineares. Os nós são dispostos

em camadas e são interligados por conexões, geralmente unidirecionais. Tais conexões são associadas a pesos que ponderam a entrada recebida por cada neurônio da rede. Nas Redes Neurais Artificiais a solução de problemas passa por uma fase inicial de aprendizagem e extração de características que são posteriormente usadas para gerar respostas ao problema (BRAGA, CARVALHO e LUDERMIR, 2000).

As Redes Neurais Artificiais têm atraído nas últimas décadas crescente atenção para previsão de séries temporais e tem conseguido aplicações de sucesso em problemas de previsão e classificação. O objetivo básico das Redes Neurais é construir modelos que simulam a inteligência do cérebro humano na máquina de modo a reconhecer padrões e regularidades nos dados de entrada, aprender com a experiência e fornecer resultados generalizados. Na previsão de séries temporais uma vantagem das Redes Neurais é a sua capacidade de modelagem não linear, sem assumir hipóteses sobre a distribuição de probabilidades dos dados que se possui (ADHIKARI e AGRAWAL, 2013). As Redes Neurais são um dos tipos mais importantes de modelos de séries temporais não paramétricos e não lineares propostos e testados para previsão de séries temporais (KHASHEY e BIJARI, 2011).

Uma rede neural pode ser pensada como um sistema que conecta um conjunto de entradas até um conjunto de saídas de uma forma possivelmente não linear. As conexões são feitas via uma ou mais camadas escondidas de neurônios. Normalmente uma função linear das entradas leva aos neurônios (nós) da camada escondida e, a estes, é aplicada uma função não linear (geralmente sigmoide ou tangente hiperbólica) para conduzir aos nós de saída (CHATFIELD, 1996).

A definição da arquitetura de uma rede é importante, pois restringe o tipo de problema que ela pode tratar. São parâmetros a considerar na arquitetura de uma rede o número de camadas, o número de nós em cada camada, o tipo de conexão entre os nós e a topologia da rede. Quanto ao tipo de conexão entre os nós a rede pode ser *feedforward* ou acíclica quando a saída de um neurônio na i -ésima camada não é usada como entrada de nós em camadas de índice menor ou igual a i e *feedback* ou cíclica, caso contrário (BRAGA, CARVALHO e LUDERMIR, 2000).

Um dos modelos de rede mais amplamente utilizados em problemas de previsão são as *feedforward* que se caracterizam por possuir três camadas (entrada, escondida e saída) ligadas por conexões sem ciclos. Os nós em cada camada são conhecidos como elementos de processamento (neurônios) (ADHIKARI e AGRAWAL, 2013). A arquitetura de uma rede neural *feedforward* com três camadas (uma de entrada, uma oculta e uma de saída) é apresentada na FIGURA 2.23.

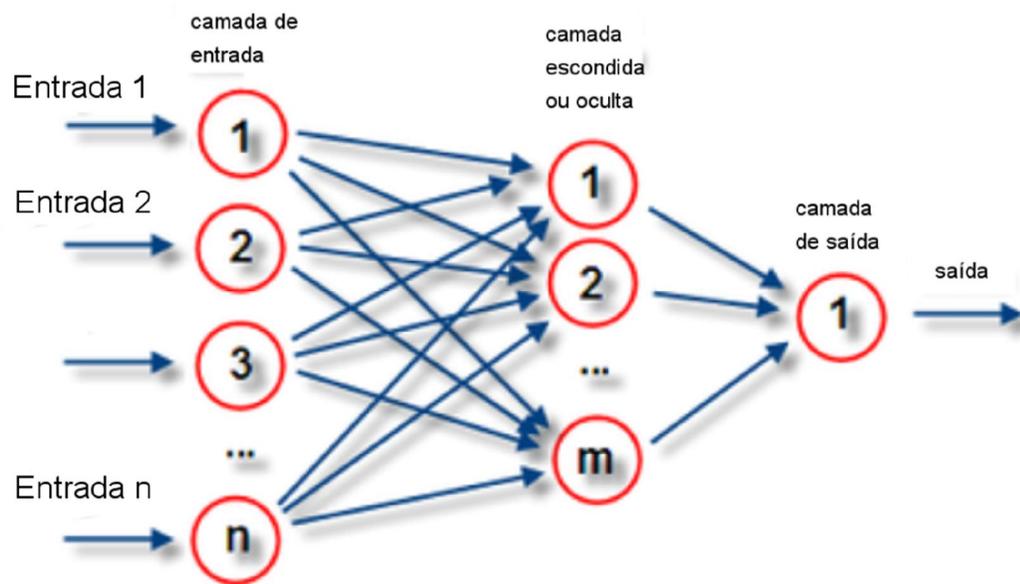


FIGURA 2.23 - Arquitetura de uma RNA *feedforward* com uma camada oculta e uma saída
FONTE: O autor

Um algoritmo de aprendizagem é um conjunto de procedimentos para adaptar os parâmetros de uma rede neural artificial para que a mesma possa 'aprender' uma função. Aprendizagem é o processo pelo qual os parâmetros de uma rede neural são ajustados e pode ser supervisionada ou não supervisionada. Na *aprendizagem supervisionada* a entrada e saída desejadas para a rede são fornecidas. O objetivo é ajustar os parâmetros da rede, de forma a encontrar uma ligação entre os pares de entrada e saída fornecidos. A rede tem sua saída calculada e comparada com a saída desejada, recebendo informações sobre o erro da resposta atual. A cada padrão de entrada submetido à rede compara-se a resposta desejada (que representa uma ação ótima para ser realizada pela rede) com a resposta calculada, ajustando-se os pesos das conexões para minimizar o erro. A soma dos erros quadráticos de todas as saídas é normalmente utilizada como medida de desempenho da rede e também como função de custo a ser minimizada pelo algoritmo de treinamento. Os exemplos mais

conhecidos de algoritmos para aprendizado supervisionado são a regra delta e a sua generalização para redes de múltiplas camadas, o algoritmo *backpropagation* (BRAGA, CARVALHO e LUDERMIR, 2000).

A saída do modelo de RNA *feedforward* é calculada através da equação

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g \left(\beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t, \quad (2.1.102)$$

onde y_{t-i} ($i = 1, 2, \dots, p$) são as p entradas e y_t é a saída, os inteiros p e q são número de nós de entrada e na camada escondida, respectivamente, α_j ($j = 1, 2, \dots, q$) e β_{ij} ($i = 0, 1, \dots, p; j = 0, 1, \dots, q$) são os pesos das conexões, ε_t é um 'choque aleatório' e α_0 e β_{0j} são os termos do *bias* (viés) e a função logística sigmoide

$$g(x) = \frac{1}{1 + e^{-x}} \quad (2.1.103)$$

é normalmente utilizada como a função de ativação não linear, dentre outras que podem ser utilizadas, como tangente hiperbólica. O modelo de rede neural nada mais é do que uma função em que as variáveis independentes são os valores passados da série temporal e a variável independente é o seu valor subsequente. Ou ainda, a rede neural é equivalente a um modelo autorregressivo não linear.

Para estimar os pesos das conexões α_j e β_{ij} procedimentos de mínimos quadrados não lineares são usados para minimizar a função erro

$$E = \sum_t e_t^2 = \sum_t (y_t - \hat{y}_t)^2, \quad (2.1.104)$$

onde \hat{y}_t é o valor previsto pela rede para a observação y_t e as técnicas de otimização usadas para minimizar a função erro são chamadas de regras de aprendizado e a mais conhecida regra de aprendizado da literatura é a *Backpropagation* (ADHIKARI e AGRAWAL, 2013).

Um modelo de rede neural deve gerar um erro pequeno não somente dentro da amostra (conjunto de treinamento), mas, também fora da amostra (conjunto de teste). Por isso se deve ter muito cuidado na escolha do número de neurônios na camada de entrada e na camada escondida. Entretanto esta é uma tarefa difícil que

não possui teoria disponível para a resposta ótima desta questão (ADHIKARI e AGRAWAL, 2013; ZHANG, 2003).

Além dos métodos de abordagem de séries temporais lineares como *ARIMA* e não lineares como *GARCH*, as Redes Neurais Artificiais (e outras técnicas de aprendizado de máquina como *Support Vector Machine*) também tem sido empregada com sucesso na modelagem de componentes não lineares da série.

2.1.8 Estatística Não Paramétrica

De uma forma geral, um procedimento não paramétrico é um procedimento estatístico que tem certas propriedades desejáveis que se mantêm sob hipóteses relativamente fracas em relação às populações subjacentes a partir do qual são obtidos os dados. Para Hollander, Wolfe e Chicken (2014, p. 1) algumas vantagens dos procedimentos estatísticos não paramétricos são:

- Exigem poucas hipóteses sobre a distribuição de probabilidade da população do qual os dados são obtidos, em particular, abandonam a hipótese tradicional de que a população subjacente é normal;
- É apenas um pouco menos eficiente do que seus ‘concorrentes’ da teoria normal quando as populações subjacentes são normais e podem ser muito mais eficientes do que os concorrentes, quando as populações subjacentes não são normais;
- São relativamente insensíveis às observações atípicas;
- Permitem abordagens em muitas situações complicadas onde métodos paramétricos são intratáveis;
- O desenvolvimento de *softwares* facilitou o cálculo rápido de valores-*p* exatos e aproximados para testes não paramétricos.

A maioria dos testes de hipóteses e intervalos de confiança abordados neste trabalho requer a hipótese de normalidade uni ou multivariada dos dados. Neste caso, estes métodos baseados em uma família de parâmetros específica (normal) são chamados de métodos paramétricos ou não livres de distribuição. Embora em casos que a distribuição subjacente seja ‘levemente’ não normal os métodos sejam

relativamente insensíveis a este desvio da normalidade, quando as distribuições subjacentes não são normais, é sugerida a utilização de métodos não paramétricos ou livres de distribuição. Estes métodos não requerem qualquer hipótese sobre a distribuição da população, somente que seja contínua. Suas vantagens vão desde a facilidade de utilização até a possibilidade de lidar com dados categóricos ou dados de postos (HINES *et al.*, 2006).

Mesmo com o desenvolvimento de métodos não paramétricos, a normalidade multivariada é importante ainda, segundo Seber (2004), pelas seguintes razões:

- Em muitas situações podem-se transformar os dados para obter normalidade;
- Alguns procedimentos baseados na normalidade multivariada são robustos aos desvios da normalidade;
- Procedimentos baseados na normalidade multivariada são frequentemente ótimos, em algum sentido, fornecendo assim um critério de comparação para procedimentos não ótimos, mas mais robustos.

O efeito da não normalidade em gráficos de controle da média e da amplitude foi investigado por Burr⁵ (1967, *apud* Ryan, 2011), que concluiu que a ausência de normalidade não é um sério problema, a menos que exista um considerável desvio da normalidade. Neste caso, é sugerido algum tipo de transformação dos dados para torná-los normalmente distribuídos e o controle da variabilidade dos dados transformados. Algumas transformações simples são dadas pelas funções *log*, raiz quadrada e recíproca, outras podem ser oriundas da família de transformações Box-Cox. Nem sempre as transformações conduzirão a normalidade, caso em que a alternativa é o método não paramétrico.

A hipótese da disponibilidade de dados que sigam uma distribuição normal multivariada garante o conhecimento da forma da distribuição da estatística T^2 (χ^2 , F ou β) e a garantia de gráficos de controle confiáveis. Entretanto, esta tarefa não é simples. No caso univariado, a hipótese do gráfico de controle de uma estatística seguir uma distribuição normal pode ser verificada por meio de *testes de adequação*

⁵ BURR, I. W. The effect of non-normality on constants for \bar{x} and R charts. **Industrial quality control**, v. 23, n. 11, p. 563-569, 1967.

ou *ajustamento*, dentre diversos disponíveis na literatura, pode-se citar os testes de Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling e Jarque-Bera.

Muitos testes têm sido propostos para avaliar a normalidade multivariada (SEBER, 2004; LOONEY, 1995; ROYSTON, 1983; BIASE e FERREIRA, 2012). Eles incluem medidas multivariadas da assimetria e da curtose, entretanto, a ausência do conhecimento da distribuição exata de certas estatísticas faz como que se obtenham apenas valores críticos aproximados para alguns procedimentos e que estes sirvam apenas como indicadores de não normalidade (MASON e YOUNG, 2002).

Looney (1995) descreve várias técnicas para avaliar a normalidade multivariada baseadas em testes bem conhecidos para avaliar normalidade univariada. Um primeiro teste seria avaliar cada variável separadamente para normalidade univariada, através de testes como assimetria, curtose e Shapiro-Wilk aplicados conjuntamente. Se a normalidade univariada for rejeitada para uma ou mais variáveis, a normalidade multivariada pode ser rejeitada. A recíproca não é verdadeira. Outras extensões multivariadas dos testes de Shapiro-Wilk, assimetria e curtose são descritas. Seber (2004, p. 149) também apresenta estimativas naturais da assimetria e da curtose para o caso multivariado que são generalizações naturais de medidas univariadas de assimetria e curtose.

Como a hipótese de normalidade multivariada garante (para a fase II, por exemplo, sem perda de generalidade) que a estatística T^2 segue uma distribuição beta, conforme o QUADRO 2.4,

$$\left[\frac{m}{(m-1)^2} \right] T^2 \sim \beta_{\frac{p}{2}, \frac{m-p-1}{2}}, \quad (2.1.105)$$

Mason e Young (2002) sugerem validar a distribuição univariada da estatística T^2 em vez da distribuição normal multivariada dos dados, realizando um teste de adequação dos seus valores para determinar se beta é a distribuição apropriada.

A ideia básica é obter, a partir de um conjunto de observações, um histograma dos valores da estatística T^2 e verificar se o mesmo se ajusta a uma distribuição apropriada (χ^2 , F ou β). Se as observações são oriundas de dados normais multivariados isto irá ocorrer, entretanto, outras distribuições multivariadas não normais poderão produzir o mesmo resultado.

Segundo Mason e Young (2002), o que estas distribuições não normais têm em comum com a normal está na distribuição amostral da estatística T^2 . A proposta sugere avaliar estimativas da curtose multivariada, além de procedimentos gráficos como o gráfico quantil x quantil ($Q - Q$) para verificar se um conjunto de dados representa uma distribuição de referência. Nele são representados graficamente os valores ordenados de T^2 contra os quantis da distribuição apropriada (χ^2 , F ou β). Forte linearidade dos pontos sugere que os valores de T^2 observados podem ser adequadamente representados pela distribuição de referência. Alerta-se ainda que a distribuição apropriada (χ^2 , F ou β) não pode ser usada indiscriminadamente para descrever a estatística T^2 no caso de não se conhecer a distribuição multivariada subjacente dos dados.

A presença de potenciais *outliers* é um problema para a avaliação da distribuição da estatística T^2 e é sugerida a sua remoção antes do ajustamento, uma vez que eles podem afetar o ajuste da distribuição apropriada correspondente, distorcer a estimativa da média e da matriz de covariância. O exame da distribuição marginal das variáveis individualmente é útil para decidir se uma distribuição, como a beta, por exemplo, está descrevendo os dados de um processo porque eles realmente seguem uma distribuição normal multivariada ou porque o ajuste produziu uma boa distribuição amostral.

Distribuições multivariadas normais implicam distribuições marginais univariadas normais, bem como qualquer subconjunto de variáveis de distribuições multivariadas normais devem ser conjuntamente normalmente distribuídas. As recíprocas não são necessariamente verdadeiras. Embora a normalidade de uma variável não garanta normalidade multivariada, ela pode melhorar a simetria global da distribuição conjunta, o que é importante quando se aplica a estatística T^2 . Caso os valores de T^2 não sigam uma distribuição apropriada, pode ser feita a tentativa de transformar os dados de cada variável individualmente para torná-la normalmente distribuída (MASON e YOUNG, 2002).

Nos casos em que uma distribuição apropriada (χ^2 , F ou β) não pode ser usada para descrever a estatística T^2 , Mason e Young (2002, p. 48) propõe três procedimentos alternativos para determinar o limite superior de controle (*LSC*) da estatística T^2 :

- Aplicar do teorema (desigualdade) de Chebyshev, que afirma que independente da distribuição de x , $P(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$, onde $k > 1$ é uma constante;
- Determinar um intervalo de confiança, livre de distribuição, para o *LSC*, usando o fato que o *LSC* representa o $(1 - \alpha)^\circ$ percentil da distribuição de T^2 . O intervalo de confiança pode ser determinado via *Bootstrap* conforme será visto mais adiante;
- Ajustar uma distribuição para a estatística T^2 usando uma técnica de *Kernel* estimativa e estimar o *LSC* pelo $(1 - \alpha)^\circ$ percentil da função distribuição ajustada para T^2 .

Alguns *softwares* como *R* e *Matlab* possuem pacotes com alguns testes para avaliar a normalidade multivariada como os testes de Royston, Henze-Zirkler e Mardia.

2.1.9 *Bootstrap*

O gráfico multivariado T^2 de Hotelling e outras formas de gráficos multivariados como o gráfico *MCUSUM* e de médias móveis exponencialmente ponderadas requerem, para sua aplicação, que os dados sejam provenientes de uma distribuição de probabilidade normal multivariada. Esta hipótese em muitas ocasiões não é verificada. Isto faz com que resultados e conclusões de dados que não a satisfazem podem não ser verdadeiros. Esta é a chamada estatística paramétrica, que presume para construção de suas teorias hipóteses sobre a distribuição de probabilidade subjacente ser normal multivariada.

Para contornar este problema, foram desenvolvidas técnicas para estabelecer limites de controle em gráficos univariados e multivariados que não dependem de uma distribuição de probabilidade específica, são os chamados gráficos de controle não paramétricos ou livres de distribuição. Sua concepção é para que sejam quase tão robustos quanto os paramétricos quando os dados têm uma determinada distribuição e que possam ser aplicáveis em situações de distribuições desconhecidas. Inicialmente será discutida uma destas técnicas, a de reamostragem *Bootstrap*.

A publicação em 1979 do primeiro artigo de Bradley Efron em *Bootstrap* foi um grande acontecimento em estatística por sintetizar algumas das ideias de reamostragem anteriores e estabelecer um novo quadro para a análise estatística baseada em simulação (DAVISON e HINKLEY, 1997).

O *Bootstrap* é uma técnica estatística recentemente desenvolvida para realizar certos tipos de inferência estatística. O desenvolvimento é recente por necessitar de computadores mais modernos para realizar com rapidez inúmeros cálculos da teoria estatística tradicional. O *Bootstrap* é um método computacional para atribuição de medidas de previsão para estimativas estatísticas (EFRON e TIBSHIRANI, 1993). O *Bootstrap* é indicado em problemas onde os procedimentos estatísticos padrões não existam ou, sejam de difícil aplicação (CHAVES NETO, 1991).

Suponha que se tenha uma amostra de dados independentes e identicamente distribuídos denotada por um vetor $x = (x_1, x_2, \dots, x_n)$ do qual se deseja calcular uma estatística de interesse $s(x)$, que pode ser, por exemplo, a média, a mediana ou um percentil da distribuição. Uma amostra *Bootstrap* $x^* = (x_1^*, x_2^*, \dots, x_n^*)$ consiste de uma amostra aleatoriamente obtida, com reposição, dos dados originais da amostra x . Por exemplo, se $n = 5$, uma amostra *Bootstrap* poderia ser $x^* = (x_3, x_5, x_1, x_5, x_3)$. Um esquema simplificado da geração de reamostras da técnica *Bootstrap* é exibido na FIGURA 2.24.

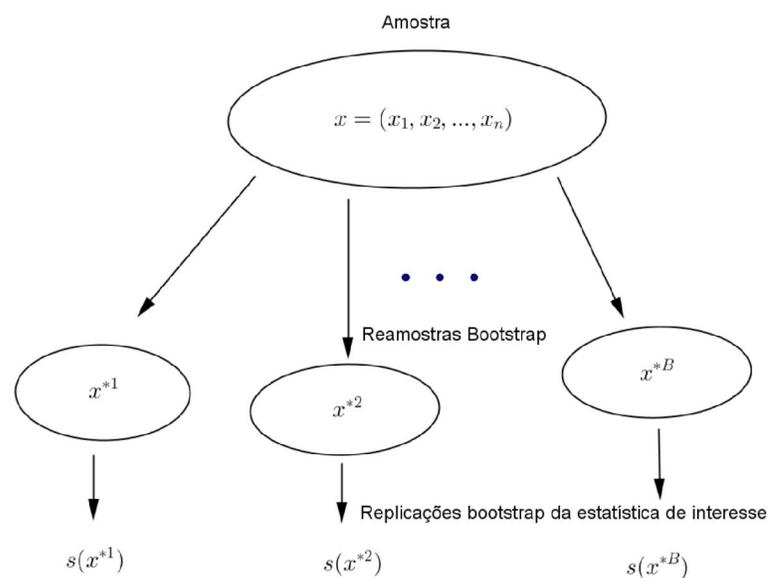


FIGURA 2.24 - Esquema da reamostragem *Bootstrap*
 FONTE: EFRON e TIBSHIRANI (1993)

Quando se tem uma amostra aleatória de tamanho n , extraída de uma distribuição de probabilidade conhecida, como a normal padrão, por exemplo, se diz que a amostra é oriunda de uma distribuição paramétrica. Porém, em muitas situações práticas, sobretudo as multivariadas, não se pode garantir ou não é verdade que a amostra seja extraída de uma distribuição conhecida.

Para Davison e Hinkley (1997), métodos que envolvem repetir um procedimento de análise de dados originais através da replicação de muitos conjuntos de dados são algumas vezes chamados de métodos computacionalmente intensivos ou *Bootstrap*, pois, usar os dados para gerar mais dados parece análogo a um truque usado por um barão fictício, que quando caiu no fundo de um lago saiu puxando a si mesmo pelos cadarços de suas próprias botas ou pelo seu próprio esforço.

Neste caso, Efron e Tibshirani (1993) sugerem que se a amostra aleatória de tamanho n é oriunda de uma distribuição de probabilidade F que gera a amostra aleatória x , isto é, $F \rightarrow (x_1, x_2, \dots, x_n)$, então se define uma função distribuição empírica \hat{F} , que atribui probabilidade $1/n$ para cada valor x_i , $i = 1, \dots, n$. A distribuição empírica pode ser simplesmente representada por um vetor de frequências $\hat{f}_k = \#\{x_i = k\}/n$. Além disso, salientam que o conjunto completo de dados da amostra $x = (x_1, x_2, \dots, x_n)$ representado reduzidamente pelo vetor de frequências, $\hat{F} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_k)$ é uma estatística suficiente para a distribuição verdadeira $F = (f_1, f_2, \dots)$, no sentido de que toda a informação de F contida em x também está contida em \hat{F} .

Dada uma amostra aleatória *Bootstrap* x^* , de tamanho n , extraída de uma distribuição de probabilidades empírica \hat{F} , que atribui probabilidade $1/n$ para cada valor observado x_i , $i = 1..n$, se $\hat{\theta}$ é uma estatística de interesse obtida da amostra x , através de alguma função $\hat{\theta} = t(\hat{F})$ que, segundo Efron e Tibshirani (1993) o princípio *plug-in* é a mesma função $t(\cdot)$ aplicada na função distribuição F que calcula o parâmetro populacional $\theta = t(F)$. Ou seja, o princípio *plug-in* estima parâmetros a partir de uma amostra.

Além de estimar algum parâmetro $\hat{\theta}$ em particular, a técnica de *Bootstrap* pode fornecer a precisão através da estimativa do erro padrão da estatística. Mais especificamente, a cada amostra *Bootstrap* corresponde uma replicação *Bootstrap* de uma estatística s , avaliada na amostra *Bootstrap* x^* , isto é, $s(x^*)$. A estimativa

Bootstrap do erro padrão é definida como o desvio padrão das replicações *Bootstrap*, assim

$$\widehat{ep}_{boot} = \sqrt{\frac{\sum_{b=1}^B [s(\mathbf{x}^{*b}) - \bar{s}]^2}{B-1}} = \widehat{ep}_B = \widehat{ep}, \quad (2.1.106)$$

onde

$$\bar{s} = \frac{\sum_{b=1}^B s(\mathbf{x}^{*b})}{B}. \quad (2.1.107)$$

Por exemplo, se $s(\mathbf{x})$ é a média amostral \bar{x} , então $s(\mathbf{x}^*) = \sum_{i=1}^n x_i^* / n$.

A técnica *Bootstrap* tem duas vantagens sobre os métodos tradicionais, a primeira que quando usada no modo não paramétrico, liberta o analista de fazer hipóteses paramétricas sobre a forma da população subjacente e, segundo, quando usada no modo paramétrico, fornece respostas mais precisas do que as fórmulas tradicionais e pode fornecer respostas para problemas em que não existam fórmulas (EFRON e TIBSHIRANI, 1993).

2.1.10 Núcleos Estimadores da Densidade (*Kernel Density Estimation*)

Quando uma amostra é extraída de uma distribuição contínua, deseja-se estimar a distribuição da população da 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 (HOLLANDER, WOLFE e CHICKEN, 2014, p. 609). Isto é particularmente útil quando a distribuição é desconhecida e requer cálculos computacionalmente intensivos.

Dada uma amostra aleatória independente e identicamente distribuída X_1, X_2, \dots, X_n de uma população com distribuição contínua, as funções densidade de probabilidade f (f.d.p.) e de distribuição acumulada F (f.d.a.) relacionam-se pela equação

$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a). \quad (2.1.108)$$

Ou ainda, para $h > 0$,

$$P\left(x - \frac{h}{2} \leq X \leq x + \frac{h}{2}\right) = \int_{x-h/2}^{x+h/2} f(t)dt = F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right). \quad (2.1.109)$$

Para f integrável e h pequeno pode-se aproximar

$$\int_{x-h/2}^{x+h/2} f(t)dt \cong hf(x). \quad (2.1.110)$$

Assim, f pode ser estimada por

$$f(x) \cong \hat{f}(x) = \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{h}. \quad (2.1.111)$$

Por sua vez, a *f.d.a.* pode ser estimada em uma amostra de tamanho n mediante a equação

$$F_n(x) = \frac{\#X_i \leq x}{n}. \quad (2.1.112)$$

Finalmente, combinando as equações (2.1.111) e (2.1.112) tem-se que

$$\begin{aligned} \hat{f}(x) &= \frac{F_n\left(x + \frac{h}{2}\right) - F_n\left(x - \frac{h}{2}\right)}{h} = \frac{\frac{\#X_i \leq x+h/2}{n} - \frac{\#X_i \leq x-h/2}{n}}{h} \\ &= \frac{\#X_i \in \left(x - \frac{h}{2}, x + \frac{h}{2}\right]}{nh}. \end{aligned} \quad (2.1.113)$$

A construção de histogramas se baseia nesta ideia, somente não exige que o centro dos intervalos seja o valor x e sim um valor predeterminado baseado no domínio dos dados que se possui e na quantidade de intervalos, que não se sobrepõe e cobrem todo domínio.

Ainda no caso específico dos histogramas, há várias regras de determinar o número de intervalos e sua largura h . A escolha de um valor maior para h pode produzir uma estimativa da densidade mais ‘suavizada’, enquanto que se h for muito pequeno pode produzir uma estimativa que é muito sensível à amostra ao invés da distribuição subjacente (HOLLANDER, WOLFE e CHICKEN, 2014, p. 613).

Embora forneça uma estimativa simples da *f.d.p.*, o histograma apresenta algumas desvantagens: não fornece uma estimativa da densidade de forma contínua, a estimativa é constante sob intervalos. As núcleo-funções (*Kernel functions*) superam

estes problemas (HOLLANDER, WOLFE e CHICKEN, 2014, p. 617; SILVERMAN, 1986).

A núcleo-estimativa da densidade de uma série $X = (X_i), i = 1, \dots, N$ no ponto x é dada por

$$f(x) = \frac{1}{Nh} \sum_{i=1}^N K \frac{(x - X_i)}{h}, \quad (2.1.114)$$

em que N é o número de observações, h é o parâmetro de suavização e K é uma núcleo-função, que 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 de núcleo-funções são

$$K_1(x) = \begin{cases} 1, & -\frac{1}{2} \leq x < \frac{1}{2} \\ 0, & \text{caso contrário,} \end{cases} \quad (2.1.115)$$

$$K_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, x \in \mathbb{R}, \quad (2.1.116)$$

$$K_3(x) = \begin{cases} \frac{3(1-x^2)}{4}, & -1 \leq x < 1 \\ 0, & \text{caso contrário,} \end{cases} \quad (\text{Kernel Epanechnikov}) \quad (2.1.117)$$

$$K_4(x) = \begin{cases} 1 - |x|, & -1 \leq x < 1 \\ 0, & \text{caso contrário.} \end{cases} \quad (2.1.118)$$

Um núcleo estimador da densidade (KDE) é uma maneira não paramétrica de estimar a função densidade de probabilidade de uma variável aleatória (WAND e JONES, 1995). A FIGURA 2.25 apresenta a função densidade de probabilidade de uma amostra aleatória de 100 pontos da distribuição normal padrão (em cinza) e 3 gráficos da densidade estimada através de um KDE com valores de $h = 0,05$ (vermelho), $h = 0,337$ (preto) e $h = 2$ (verde).

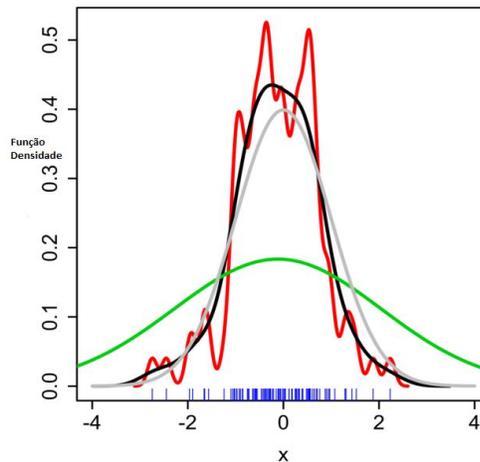


FIGURA 2.25 - Comparação de diferentes valores de h para uma amostra da distribuição normal padrão
FONTE: O autor

A técnica KDE oferece uma maneira não paramétrica de estimar a função de regressão sem a especificação de um modelo paramétrico (quando a família de funções de um modelo pode ser especificada por um número finito de parâmetros) e pode ser estendida a outros problemas mais complexos e com aplicações em diversos campos, como a medicina, engenharia e economia (WAND e JONES, 1995, p. xi).

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 pode não ser adequado 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”. Modelos paramétricos e não paramétricos não devem ser considerados competidores. Em alguns casos, modelos não paramétricos podem sugerir o uso de modelos paramétricos mais simples e em outros, modelos paramétricos podem se mostrar bastante inadequados e sugerir os modelos não paramétricos (WAND e JONES, 1995, p. 3).

Um problema de regressão muito comum é, a partir dos dados de amostra aleatória estimar a função densidade de probabilidade (*f.d.p.*). 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 para a *f.d.p.* O histograma é um exemplo de estimativa da *f.d.p.* não paramétrica (WAND e JONES, 1995, p. 5).

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. A FIGURA 2.26 exibe o efeito do núcleo estimador da densidade (KDE) sobre uma amostra de 1000 observações de uma combinação de distribuições normais. O gráfico sólido indica a densidade estimada e o gráfico pontilhado indica a verdadeira densidade. Os valores de h são: em (a) $h=0,06$, em (b) $h=0,54$ e em (c) $h=0,18$.

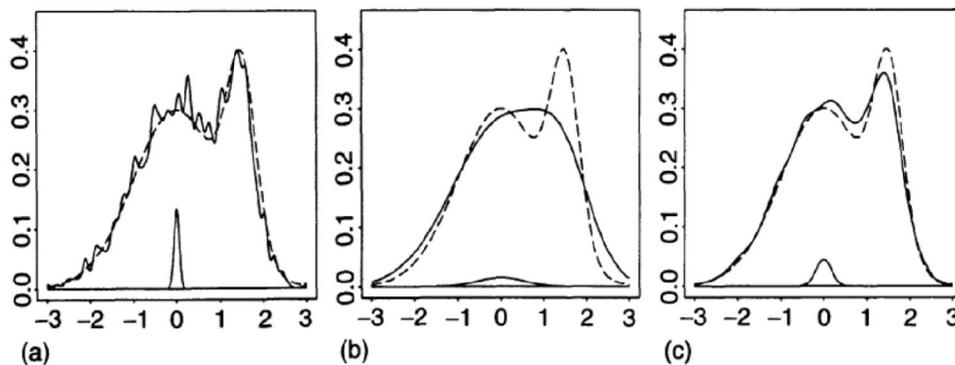


FIGURA 2.26 - Efeito de diferentes valores de h no KDE
 FONTE: WAND e JONES, 1995, p. 13

O desempenho do núcleo estimador da densidade pode ser medido por uma generalização funcional do MSE (erro quadrático médio) que mede quão perto uma estimativa $\hat{\theta}$ está do parâmetro alvo θ , dado por

$$MSE = E(\hat{\theta} - \theta)^2 = \underbrace{V(\hat{\theta})}_{\text{variância}} + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{viés}^2}. \quad (2.1.119)$$

Que toma a forma MISE (erro quadrático integrado médio) dada por

$$MISE\{\hat{f}(\cdot, h)\} = E\left(\int [\hat{f}(x, h) - f(x)]^2 dx\right). \quad (2.1.120)$$

Na sua forma assintótica (grandes amostras) *AMISE*, é possível mostrar que depende de h de uma forma mais simples que facilita a interpretação de h no desempenho do núcleo estimador da densidade em que uma parcela de *AMISE* (do viés ao quadrado integrado) é assintoticamente proporcional a h^4 (para esta quantidade decrescer deve-se tomar h pequeno) e a outra parcela da variância integrada é proporcional a nh^{-1} (portanto cresce quando h decresce). Esta ‘negociação’ entre variância e viés é a regra crítica que determina o valor ótimo de h .

Mais especificamente, assintoticamente, segundo Hollander, Wolfe e Chicken (2014, p. 623-627), para uma núcleo-função K

$$AMISE = \frac{\int K^2}{nh} + \frac{\sigma_K^4 h^4 \int (f'')^2}{4}, \quad (2.1.121)$$

onde $\sigma_K^2 = \int x^2 K(x) dx$. Minimizando a equação (2.1.121) em relação ao núcleo K produz o núcleo de Epanechnikov, já, se for minimizada quanto à h , o valor ótimo de h que minimiza $AMISE$, usando o núcleo da normal e assumindo que os dados vem de uma população normal com média 0 e variância σ^2 , é

$$h = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \cong 1,06 \cdot \sigma \cdot n^{-1/5}. \quad (2.1.122)$$

2.1.11 Problema de Programação Matemática

Um modelo de otimização contínua é resumidamente composto de variáveis, denominadas variáveis de decisão; uma função das variáveis de decisão, chamada objetivo, a ser otimizada (maximizada ou minimizada) e um conjunto de restrições que são equações ou inequações envolvendo as variáveis de decisão a serem atendidas por eventuais soluções viáveis. A pesquisa operacional consiste no desenvolvimento de técnicas e métodos para resolução de problemas e tomada de decisões. A programação matemática colabora de várias formas com a pesquisa operacional na modelagem e solução de problemas de decisão (GOLDBARG e LUNA, 2000).

Os principais modelos de pesquisa operacional são chamados de programação (planejamento) matemática e são importantes modelos quantitativos para solução de problemas de otimização. As técnicas de solução para tais tipos de problemas podem variar de acordo a classificação de problemas de programação matemática: lineares (aqueles em que a função objetivo e as restrições são todas funções lineares das variáveis de decisão), não lineares (com algum tipo de não linearidade na função objetivo ou restrições) e de programação inteira (com a restrição de que alguma variável seja um número inteiro). Outras famílias de problemas decorrem daí, como problemas de conexão, fluxo em redes, caixeiro viajante e de roteamento, bem como estratégias heurísticas de resolução (GOLDBARG e LUNA, 2000).

Para maiores detalhes sobre os modelos de programação matemática e algoritmos de solução recomendam-se as seguintes referências: Yang(2008), Murty (1983), Ragsdale (2012) e Tawarmalani e Sahinidis (2002).

2.2 REVISÃO DE LITERATURA

Nesta seção é realizada uma revisão da literatura no que tange aos artigos científicos que tratam dos métodos e problemas similares ao que se tratou neste trabalho. Embora não seja possível muitas vezes fazer uma completa separação, a revisão dos artigos foi separada em cinco subseções para facilitar a leitura.

2.2.1 Gráficos Multivariados, Experiências Internacionais em Barragens

Houve um considerável entusiasmo no controle estatístico de qualidade, durante e após a segunda guerra mundial, principalmente pelo fato de exigências de controle de qualidade em contratos dos governos americano e inglês e também na retomada da produção não militar pós-guerra. Depois de um período de estagnação nos anos 60 pela falta de fontes computacionais adequadas, o uso de técnicas de controle de qualidade multivariadas ressurgiu com o desenvolvimento de computadores mais poderosos e acessíveis a toda sociedade (JACKSON, 1985).

Atualmente existem muitas situações nas quais é importante monitorar duas ou mais características de qualidade relacionadas. O monitoramento independente pode levar a interpretações equivocadas. Processos em que há variáveis altamente correlacionadas podem dar sinais de fora de controle incorretos e enfraquecer a eficiência do gráfico. Problemas de monitoramento de diversas características de qualidade simultaneamente estão na área de controle estatístico de qualidade multivariada e a principal ferramenta para sua implementação são os gráficos de controle (BERSIMIS, PSARAKIS e PANARETOS, 2007, p. 517).

Para Mastrangelo, Runger e Montgomery (1996), as técnicas de gráficos de controle multivariadas têm várias vantagens sobre os gráficos univariados. O comprimento médio da sequência (CMS) ou a taxa de erro tipo I pode ser preservado, relações entre as variáveis podem ser explicadas e uma única resposta para a questão “O processo está sob controle?” pode ser dada.

Os gráficos para o controle da média de um processo são geralmente baseados na distância estatística de *Mahalanobis* e esta estatística (distância) no estabelecimento da fase I resume-se em dois casos. No primeiro, quando se dispõe de m amostras de tamanho n , denotadas por X_i , para $i = 1, 2, \dots, m$, provenientes de um vetor p -variado de média μ e covariância Σ conhecida, são avaliadas as estatísticas

$$\chi_i^2 = n(X_i - \mu)' \Sigma^{-1} (X_i - \mu). \quad (2.2.1)$$

No segundo caso, se a média e a matriz de covariância não são conhecidas e são estimadas das amostras por \bar{X} e S , tem-se a estatística

$$T_i^2 = n(X_i - \bar{X})' S^{-1} (X_i - \bar{X}). \quad (2.2.2)$$

As estatísticas χ_i^2 e T_i^2 , sob a hipótese de amostras independentes e identicamente distribuídas de dados oriundos de distribuição multivariada de p variáveis, representam a distância ponderada de qualquer ponto do alvo, que no caso, é a média do processo (BERSIMIS, PANARETOS e PSARAKIS, 2005, p. 2).

Cheng e Zheng (2013) criaram dois modelos para relacionar os efeitos entre leituras de instrumentos (piezômetros e tensômetros) e fatores ambientais (nível do reservatório, temperatura do ar e quantidade de chuva) na barragem da usina de Mianhuatan, na China. Um modelo tinha o objetivo de extrair o efeito das variáveis ambientais através da análise da matriz de covariância dos instrumentos e das componentes principais e a interpretação das componentes principais como as variáveis ambientais. Com base na diferença entre os valores da instrumentação e de um modelo de regressão aplicada nas componentes principais se estabeleciam gráficos de controle dos resíduos cuja distribuição de probabilidade é calculada baseada no método KDE (*Kernel density estimation*), que não requer a hipótese de normalidade.

O outro método utiliza as variáveis ambientais (causa) e as leituras de instrumentação de vários instrumentos do mesmo tipo (efeito), extrai variáveis latentes do conjunto de instrumentos e cria modelos de Redes Neurais e de *least squares support vector machine* (LS-SVM) entre as variáveis causa (entrada) e as componentes principais (saída) que, depois de treinadas, servem para construir e avaliar gráficos de controle das variáveis latentes. Os métodos mostraram boa capacidade de reduzir a taxa de falsos alarmes e detectar instrumentos defeituosos.

O diagnóstico de valores singulares (pontos fora de controle) foi proposto no monitoramento de segurança de barragem, com um estudo de caso na usina hidrelétrica localizada na província de Fujian na China, via análise de componentes principais e gráfico de controle multivariado T^2 de Hotelling. Os valores do monitoramento da segurança da barragem podem indicar mudanças estruturais da barragem ou serem reflexo de alterações ambientais. Em geral, o monitoramento destes 'erros' (ou valores singulares) em um sistema de monitoramento de barragem, pode ser dividido em erro sistemático e erro aleatório. A metodologia mostrou eficiência na redução da taxa de falsos alarmes (GU *et al.*, 2011).

Mata, Castro e Costa (2014) propõe um modelo estatístico HTT (hidrostático, térmico, temporal) para interpretar os deslocamentos de uma barragem de concreto (em Alto-Lindoso, Portugal) em substituição a função sazonal de modelos hidrostático sazonal no tempo (HST) usando as temperaturas registadas para representar melhor o efeito sobre o comportamento térmico da barragem. As metodologias propostas na construção de modelos estatísticos HTT baseiam-se no uso de análise de componentes principais (ACP) aplicados às temperaturas registadas no corpo da barragem de concreto. A ACP é usada para escolha dos termômetros na construção do modelo HTT e para representar o efeito térmico no modelo. Segundo os autores, a vantagem deste método é que o efeito térmico é representado pela temperatura real medida no corpo da barragem de concreto. (MATA, CASTRO e COSTA, 2014)

Yu *et al.*(2010) aplicaram aos medidores de junta na Chencun *Hydropower Station* na China um modelo que busca extrair as componentes principais dos dados da instrumentação e estabelecer um modelo hidrostático sazonal no tempo (HST) entre as variáveis: 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. O trabalho resultou em redução de dados redundantes que diminui o atraso entre a leitura e análise dos dados de instrumentação, separação do ruído (devido às causas aleatórias) e a redução de falsos alarmes.

Foram desenvolvidos dois métodos aplicados à obtenção de tendências da deformação estática na barragem da usina de Fei-Tsui localizada em Taiwan. Um método utiliza a análise de uma série temporal baseada em princípios de estatística multivariada não paramétrica chamada análise espectral singular combinada com um modelo autorregressivo. O segundo combina análise de componentes principais não lineares com uma rede neural autoassociativa. Ambos têm o objetivo de definir alertas precoces sobre mudanças no comportamento da barragem com relação à deformação. Foram usados dados de pêndulos horizontais e verticais. Dados da medida de temperatura diária também foram usados na tentativa de extrair o complexo comportamento não linear causado pela variação das condições ambientais da variação das características medidas (LOH, CHEN e HSU, 2011).

Xu *et al.* (2012) monitoraram a componente do deslocamento térmico em uma usina hidrelétrica localizada na província de Fujian, na China, por meio de três modelos de regressão que decompõe o deslocamento em três parcelas, uma devida a uma componente hidrostática (polinomial), outra componente térmica (trigonométrica e exponencial) e a outra uma componente devida o efeito do tempo (logarítmica). O objetivo é melhorar um modelo tradicional de decomposição em componentes hidrostática, sazonal e do tempo. Os modelos variam de acordo com a forma da função que descreve cada componente. Os pontos de referência para construção dos modelos de deslocamentos são um conjunto de pêndulos diretos e invertidos instalados na barragem para monitorar o deslocamento horizontal ao longo e através (longitudinal e transversal) da direção do rio. Para cada instrumento foram comparados os valores medidos com os modelos ajustados através do coeficiente de correlação e do desvio padrão.

Quintanilha *et al.*, (2013) monitoraram a movimentação de uma barragem (não especificada no artigo) por meio da auscultação geodésica de medidas de distância, em várias épocas, de posições consideradas fixas, os pilares, a outras

posições, os alvos, com o objetivo de identificar os alvos que apresentam medidas atípicas de modo a fim indicar ações corretivas. A metodologia consistiu da construção de gráfico de controle multivariado T^2 aplicado aos resíduos de um modelo de mínimos quadrados ajustado aos dados geodésicos, os quais se espera comportamento aleatório.

Motivados pela crescente quantidade de acidentes em barragens, seja por envelhecimento, terremotos ou mudanças não convencionais das condições do tempo, Jeon *et al.* (2009) desenvolveram em conjunto com a sociedade coreana de recursos hídricos um sistema de gerenciamento de segurança de barragens. Foi criado um sistema de monitoramento e instrumentação automatizado que recebe dados de várias barragens numa central de base de dados, se limites pré-estabelecidos forem ultrapassados, um alerta precoce é dado aos gerenciadores da barragem. Periodicamente é checado se há mau funcionamento de instrumentos e a confiabilidade dos dados medidos. Gráficos envolvendo possíveis causas e efeitos também são muito úteis.

2.2.2 Algumas Experiências com ACP

Segundo Kourti (2005, p. 214), cresceu nos últimos 15 anos o interesse industrial pelo monitoramento multivariado na detecção de falhas. Ao detectar alguma anormalidade no uso de gráficos univariados para monitorar cada variável de qualidade de um produto, normalmente várias variáveis apresentam-se fora de suas regiões de controle, devido à correlação existente entre elas, e torna-se difícil determinar a fonte do problema. O uso do método de variáveis latentes, para monitorar globalmente, diminui o número de gráficos a serem analisados, facilita o entendimento e permite detectar a fonte do problema com mais antecedência que os gráficos univariados devido à natureza não correlacionada das variáveis latentes.

Miletic *et al.*, (2004, p. 822) discutem a aplicação de análise de componentes principais e mínimos quadrados parciais para explorar uma infraestrutura de dados disponível de modo a obter retorno de investimentos no ambiente industrial, seja

através de modelos de previsão, detecção de falhas e controle de qualidade do processo.

Kourti (2005, p. 217) relata que quando o número de variáveis que estão sendo monitoradas é muito grande, pode ocorrer problemas no cálculo da estatística T^2 de Hotelling devido à singularidade. Este problema pode ser superado pelo uso das k primeiras variáveis latentes que modelam a maior parte da variabilidade do processo.

O método de projeção no espaço de variáveis latentes leva ao gráfico T^2 de Hotelling das k primeiras variáveis latentes. As novas variáveis t_i podem ser os escores de um modelo de análise de componentes principais sobre as p variáveis de entrada do processo X (a matriz $n \times p$ das medidas do conjunto de dados históricos) ou os escores de um modelo de mínimos quadrados parciais entre as variáveis de entrada e as variáveis de qualidade Y (uma matriz $n \times m$). Neste caso a estatística torna-se

$$T_k^2 = \sum_{i=1}^k \frac{t_i^2}{\lambda_i} = \sum_{i=1}^k \frac{t_i^2}{s_{t_i}^2}, \quad (2.2.3)$$

onde $s_{t_i}^2$ é a variância estimada da variável latente t_i . O gráfico verifica se uma nova observação X das p variáveis de entrada do processo projetada sobre o hiperplano de componentes principais fica dentro dos limites determinados pelos dados de referência. O limite superior de controle (LSC) da estatística T_k^2 baseado nas k primeiras componentes principais, obtido a partir da distribuição- F é

$$LSC = \frac{(n^2 - 1)k}{n(n - k)} F_{\alpha, k, n-k}. \quad (2.2.4)$$

Também é comum estudar o gráfico de controle do erro de previsão quadrático (*square prediction error*, em inglês, SPE) que calcula a soma dos quadrados dos desvios de uma nova observação X em relação ao previsto pelo modelo de referência \hat{X} , isto é, a estatística

$$SPE = \sum_{i=1}^p (X_i - \hat{X}_i)^2, \quad (2.2.5)$$

cujos limites superior de controle (LSC), determinado por Jackson e Mudholkar (1979, p. 342), é

$$LSC = Q_\alpha = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{1/h_0}, \quad (2.2.6)$$

onde z_α é valor que corresponde ao $100(1 - \alpha)$ percentil da distribuição normal padrão, α a probabilidade de erro tipo I e

$$\theta_i = \sum_{j=k+1}^p \lambda_j^i = tr(E^i), \quad (2.2.7)$$

para $i = 1, 2, 3$, λ_i é o i -ésimo autovalor da matriz de covariância e

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}. \quad (2.2.8)$$

Deve-se enfatizar que este limite superior de controle presume a hipótese de normalidade multivariada dos dados.

Cheng e Zheng (2013) utilizam o gráfico de controle do *SPE*, ao considerar que o resíduo é o resultado da remoção do efeito das variáveis ambientais no monitoramento dos dados e que quando a barragem estiver operando de forma normal, o valor da norma do *SPE* deve permanecer sob determinados valores de controle. Os autores supõem ainda que quando existirem anormalidades na estrutura da barragem ou mau funcionamento de instrumentos, a influência de outros fatores crescerá significativamente e a norma do *SPE* excederá os limites de controle. Entretanto, a distribuição do *SPE* não pode ser assumida ser normalmente distribuída e estimam o *LSC* por meio do método *KDE*.

Quando as componentes principais são utilizadas como técnica de redução de dados, ferramenta de diagnóstico ou controle, os resíduos associados às componentes principais são úteis para verificar o ajuste ao modelo e servir de teste para detecção de *outliers* (JACKSON e MUDHOLKAR, 1979, p. 341).

Na utilização de componentes principais, é sugerido primeiramente que seja feito o teste da soma dos quadrados dos resíduos (teste Q ou *SPE*), pois se este apresentar significância, ele lança dúvida sobre outros resultados, no sentido de usualmente estar conectado com um teste de medida de algum tipo de discrepância como dados registrados incorretamente. Caso os resíduos não sejam estatisticamente significativos, deve-se avaliar a estatística T^2 , se esta também não for significativa, o

processo é considerado em estado normal, caso contrário deve-se procurar as causas nas componentes principais (JACKSON, 1985).

Para Jackson (1980, p. 212), um valor significativo da estatística Q se deve a presença de outra fonte de variabilidade real não explicada pelas componentes principais ou o excesso de variabilidade aleatória, sobretudo esta última.

O uso de modelos de gráficos de controle multivariados como *MCUSUM* e *MEWMA* pode ser impraticável quando há um número elevado de variáveis com colinearidades. Um procedimento usual para reduzir a dimensão do espaço de variáveis é usar métodos de projeção tais como *ACP* e *PLS*. Estes métodos buscam construir modelos a partir de um conjunto histórico de dados que é assumido estar sob controle (BERSIMIS, PSARAKIS e PANARETOS, 2007, p. 528).

A FIGURA 2.27 exhibe um esquema de um procedimento de controle estatístico de processo multivariado. Dados de uma rotina operacional de boa qualidade (fase I) constituem o conjunto de treinamento com as variáveis causa (entrada) e as variáveis de qualidade (saída) do processo. Um modelo é desenvolvido, seja de componentes principais ou mínimos quadrados, por exemplo. Valores de controle para os gráficos T^2 e do *SPE* são obtidos e novos dados da operação *on-line* (fase II) são testados nas estatísticas de controle.

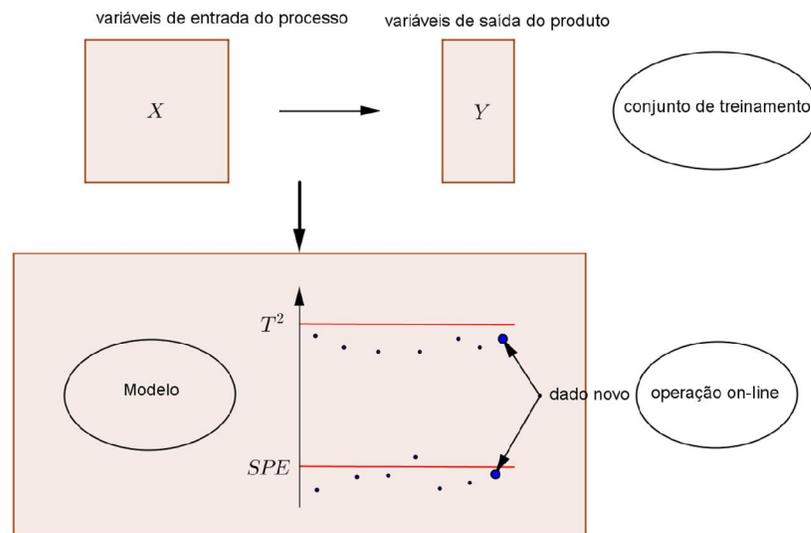


FIGURA 2.27 - Esquema de procedimento de controle estatístico multivariado
FONTE: Adaptado de Kourtí (2005, p. 219)

Quando dados são sinalizados como fora dos limites de controle de um gráfico de controle multivariado, Kourti (2005, p. 221) sugere, com cautela, a utilização de procedimentos adicionais de investigação como os gráficos de contribuição. Por exemplo, na contribuição para o gráfico do erro quadrático de previsão, para cada variável i do conjunto original pode-se simplesmente calcular a contribuição devida a ela, dada por $(x_i - \hat{x}_i)^2$.

No âmbito do controle estatístico de qualidade, os métodos na fase I são comparados avaliando a probabilidade de decidir se o processo está fora de controle, ou seja, instável. Na Fase II, a ênfase está em detectar tendências ou mudanças de processo o mais rápido possível. Isto é geralmente medido por parâmetros da distribuição do comprimento da sequência, onde o comprimento médio da sequência (CMS ou ARL) é o número de amostras tomadas antes de um sinal de fora do controle ser emitido. O CMS (ou ARL) é frequentemente usado para comparar o desempenho de métodos de gráfico de controle concorrentes na Fase II, geralmente sob a hipótese de tendências ou deslocamentos nos parâmetros do modelo estimado a partir da Fase I (WOODALL *et al.*, 2004, p. 311).

Uma análise de forma multivariada da instrumentação de uma barragem utilizou-se de regras para estabelecer quando um sinal de fora de controle é devido a alguma razão sazonal, mau funcionamento de instrumento ou indicação de mudança no comportamento da barragem (CHENG e ZHENG, 2013).

Quando se busca diagnosticar causas para algum sinal de fora de controle num gráfico multivariado é sugerida a utilização de gráficos de contribuição de cada variável. Variáveis com alta taxa de contribuição devem ser investigadas, se for o caso, com gráficos univariados. Não obstante, deve haver cautela na interpretação dos gráficos de contribuição. As variáveis altamente correlacionadas com variáveis com maior contribuição para sinais de fora de controle devem também ser levadas em conta na busca por uma causa atribuível (KOURTI, 2005, p. 214).

Miletic *et al.* (2004, p. 823) afirmam que em modelos de aplicação de controle estatístico de qualidade no ambiente industrial, a matriz de covariância amostral (e sua inversa) é frequentemente a base das estatísticas de interesse. Quando modelos de previsão lineares são desenvolvidos é comum se trabalhar com grande quantidade

de dados e variáveis que são altamente correlacionadas, resultando na problemática de matrizes de covariância mal condicionadas. Isto pode ser superado por métodos de projeção como a análise de componentes principais.

2.2.3 Gráficos de Controle Não Paramétricos

A hipótese de normalidade pode não ser realmente verdadeira. Neste caso, Boone e Chakraborti (2012) afirmam que o desempenho do gráfico de controle cai tanto na taxa de falsos alarmes quanto na habilidade de detectar mudanças na característica de qualidade que está sendo avaliada e propõe a utilização de procedimentos não paramétricos ou livres de distribuição, estes têm a vantagem de não requerer nenhuma hipótese sobre a distribuição dos parâmetros.

O método não paramétrico *Bootstrap* pode ser usado para estimar a distribuição amostral de uma estatística usando somente a hipótese de que a amostra seja representativa da população do qual é extraída e que as observações sejam independentes e identicamente distribuídas (NIAKI e ABBASI, 2007).

Niaki e Abbasi (2007) desenvolveram uma metodologia para obter limites de controle para atributos (tipo defeituoso e não defeituoso) baseada no método *Bootstrap* em que intervalos de confiança simultâneos para os atributos são construídos. Baseado em dados multivariados sob controle submetidos a mudanças no vetor de médias e simulações com dados seguindo a distribuição de Poisson, o desempenho do método proposto foi comparado com procedimentos tradicionais e mostrou-se melhor tanto para o comprimento médio da sequência sob controle (ARL_0) e fora de controle (ARL_1).

Martin e Morris (1996, p. 350) foram um dos primeiros a introduzir métodos para gráficos de controle multivariados não paramétricos. Este trabalho propôs uma nova abordagem para construir limites de controle com base na densidade dos dados. A metodologia combinou as técnicas *Bootstrap* padrão e de núcleo estimadores da densidade (KDE) para produzir uma nova estatística não paramétrica para calcular limites de confiança. A nova estatística foi comparada com a do gráfico T^2 de componentes principais obtidas pelos métodos *ACP* e *PLS* em um processo aplicado.

Na avaliação da região sob controle da fase I, a nova estatística mostrou reduzir significativamente a dispersão dos dados, sobretudo na periferia do conjunto, comparativamente ao gráfico T^2 no caso do método ACP e na detecção de mudanças na fase II, no caso do método PLS, a nova estatística permitiu a identificação mais rápida de problemas na qualidade final de um produto.

Phaladiganon *et al.*(2011) propuseram um gráfico de controle multivariado T^2 baseado na técnica *Bootstrap* para monitorar um processo quando a distribuição dos dados não é normal ou é desconhecida. Realizaram comparações do desempenho do gráfico proposto com o gráfico T^2 de Hotelling que presume normalidade multivariada e com um gráfico de controle não paramétrico baseado na técnica *KDE*. Os resultados mostraram que o gráfico baseado na técnica *Bootstrap* teve melhor desempenho que o gráfico de Hotelling e desempenho comparável com o gráfico do método *KDE* para situações de desvio da normalidade e desempenho comparável ao do gráfico T^2 de Hotelling e do método *KDE* em termos do ARL_0 na situação da normalidade dos dados.

Phaladiganon *et al.*(2013) enfatizam que a análise de componentes principais tem sido utilizada no monitoramento de processos industriais caracterizados por um grande número de características de qualidade de interesse. A correlação existente entre as variáveis, geradora de falsos alarmes, pode ser superada com o uso de componentes principais não correlacionadas. Entretanto, o estabelecimento de limites de controle para os gráficos tradicionais T^2 e Q requer a hipótese de normalidade, desta forma, os autores propõem o uso da técnica não paramétrica *KDE* (*Kernel Density Estimation*) aliada à amostragem de *Bootstrap* para estabelecer limites de controle para estes gráficos. É salientado que usar somente as k componentes principais, pode não permitir a detecção de uma observação fora de controle causada por uma das $p - k$ componentes principais. Portanto, o uso do gráfico Q é necessário para detectar mudanças devido às outras componentes principais.

Recentemente, Phaladiganon *et al.*(2013) desenvolveram gráficos de controle não paramétricos das componentes principais, que transformam um conjunto de variáveis correlacionadas num conjunto de variáveis não correlacionadas com o objetivo de reduzir falsos alarmes em gráficos de controle multivariados. Como a hipótese de normalidade multivariada é exigida para obtenção de limites de controle para os gráficos, o uso de técnicas não paramétricas como *Bootstrapping* e *Kernel*

Density Estimation é proposta para estabelecer limites de controle e comparações de desempenho de gráficos através de simulações.

Cheng e Zheng (2013) desenvolveram modelos de gráficos de controle multivariados aplicados a instrumentação de barragem de usina hidrelétrica em que usam técnicas não paramétricas como *Kernel Density Estimation* e *Least Squares Support Vector Machine* para estabelecer os limites de controle.

Boone (2010) avaliou o desempenho do gráfico Z, proposto por Kalgonda e Kulkarni (2004), para tratar com dados multivariados autocorrelacionados e comparou com o gráfico de Hotelling e gráficos multivariados de médias móveis exponencialmente ponderadas através de simulações. As simulações também procuraram avaliar o desempenho do gráfico Z no estudo do efeito da estimação de parâmetros e da não normalidade usando distribuições multivariadas t e gama e propôs quatro gráficos multivariados não paramétricos. O autor fez uma aplicação com dados de uma pequena usina hidrelétrica onde a taxa de fluxo da água que passa por um conduto forçado e a energia gerada a partir do gerador (em quilowatts) são monitorados. O gráfico multivariado Z assume normalidade dos dados.

Chakraborti (2011) aponta que os métodos de controle estatístico de qualidade têm colaborado na melhoria da qualidade de produtos no sentido de identificar e reduzir causas atribuíveis de variabilidade que possam estar presentes, neste caso, as causas aleatórias de variabilidade são frequentemente assumidas seguirem uma distribuição paramétrica, geralmente a normal, este tipo de controle estatístico de qualidade é chamado de paramétrico. A escolha da estatística a ser analisada e os cálculos dos limites de controle usualmente dependem da forma da distribuição de probabilidade subjacente e as propriedades dos gráficos de controle podem deixar de ser verdadeiras sem a hipótese da distribuição paramétrica subjacente.

Chakraborti (2011) afirma ainda que, embora o grande uso do controle estatístico de qualidade paramétrico, há dúvidas sobre o desempenho de seus métodos nas ocasiões onde existe incerteza sobre a forma da distribuição de probabilidade, isto suscita a necessidade de gráficos de controle mais flexíveis e robustos que não exijam a hipótese de normalidade ou qualquer outra hipótese sobre a distribuição. Os gráficos de controle não paramétricos (ou livres de distribuição) podem oferecer alternativas úteis aos gráficos de controle paramétricos. Embora não

haja muitos testes não paramétricos, podem ser desenvolvidos testes de hipóteses e intervalos de confiança para os testes não paramétricos com desempenho comparado a sua contrapartida paramétrica mesmo no caso da distribuição subjacente ser normal. Enfatiza que na prática, raramente se pode dispor de dados normais e, portanto, as técnicas não paramétricas de controle estatístico de qualidade são aconselhadas.

Chakraborti, Van Der Laan e Bakir (2001) apontam as seguintes vantagens nos gráficos de controle não paramétricos: simplicidade, não necessitam assumir uma distribuição particular do processo, o comprimento da sequência sob controle é o mesmo para todas distribuições contínuas, bem como a taxa de falsos alarmes, o que facilita a comparação de diferentes gráficos de controle, mais robusto e menos afetado por *outliers*, é mais eficiente na detecção de mudanças quando a verdadeira distribuição é significativamente não normal.

Quando o controle de qualidade multivariado apresenta autocorrelação sugere-se a construção de modelos de vetor autorregressivo (*VAR*), que é uma generalização dos modelos *ARIMA* univariados. O monitoramento dos resíduos de séries temporais univariados tem o objetivo de separar na série os padrões sistemáticos e expor causas de mudanças, neste sentido se avalia um gráfico de causas comuns de variação e outro de causas especiais, onde as causas comuns monitoram a variação sistemática representada por um modelo *ARIMA* e as causas especiais nos resíduos do modelo *ARIMA*. Mudanças significativas no processo, seja na média, na variância ou nos parâmetros do modelo *ARIMA*, refletem-se no gráfico de controle dos resíduos. O autor sugere dispensar a parte de médias móveis do modelo porque no caso multivariado esta é, em certo sentido, transferida para a parte autorregressiva, mesmo que de uma ordem maior. Após o ajustamento de um modelo *VAR*, o vetor resíduo num instante de tempo t , que é a diferença entre o valor observado num conjunto de dados histórico e o valor previsto pelo modelo no instante de tempo $t-1$ para t , sob algumas hipóteses, pode ser modelado pelo gráfico T^2 (PAN e JARRETT, 2012).

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. É usada uma técnica de núcleo estimadores da densidade (*Kernel Density Estimation*)

para a distribuição da estatística T^2 e do limite superior de controle quando os dados não são normais multivariados (CHOU, MASON e YOUNG, 2001).

Vermaat *et al.*(2003) realizaram uma comparação entre gráficos de controle para observações individuais (amostras de tamanho $n = 1$) de gráficos de Shewhart das médias móveis com gráficos não paramétricos baseados em quantis de distribuição *Bootstrap* e núcleo-estimadores (*Kernel estimators*) em diversos cenários em estudos de simulação com dados de diversas distribuições paramétricas na fase II do gráfico de controle. Seus resultados mostraram que os quantis empíricos obtidos da técnica *Bootstrap* tiveram excelente desempenho em todas as distribuições consideradas, mostrando-se robusta a desvios da normalidade e com bom desempenho sob esta hipótese. O comportamento dos gráficos de controle obtidos pelos núcleo-estimadores foi bastante similar aos gráficos obtidos por *Bootstrap* com exceção quando os dados são da distribuição uniforme.

Liang (2005), com o objetivo de superar as limitações de métodos para controle de um esquema de monitoramento multivariado que dependem da hipótese de normalidade e melhorar o desempenho de processo de monitoramento em tempo real, fez um estudo de caso industrial em que combinou o uso de análise de componentes principais para definição dos limites de controle com uma núcleo- função de estimativa da densidade de probabilidade multivariada e seus resultados mostraram as vantagens desta abordagem em relação ao método convencional de assumir a hipótese de normalidade: diminuição do número de falsos alarmes, identificação de comportamento anormal com antecedência e redução da variabilidade dos dados. Foram retidas quatro componentes principais de um total de catorze variáveis e avaliadas as estatísticas *SPE* e T^2 .

Uma evolução das técnicas estatísticas multivariadas de controle de processos (bivariado, sem perda de generalidade) é apresentada na FIGURA 2.28. No caso a) duas variáveis monitoradas separadamente por gráficos de Shewhart, no caso b) variáveis com distribuição normal bivariada e independentes, no caso c) variáveis com distribuição normal, mas correlacionadas, e no caso d) onde os limites de controle do gráfico não dependem da distribuição e 'aprendem' com os dados disponíveis.

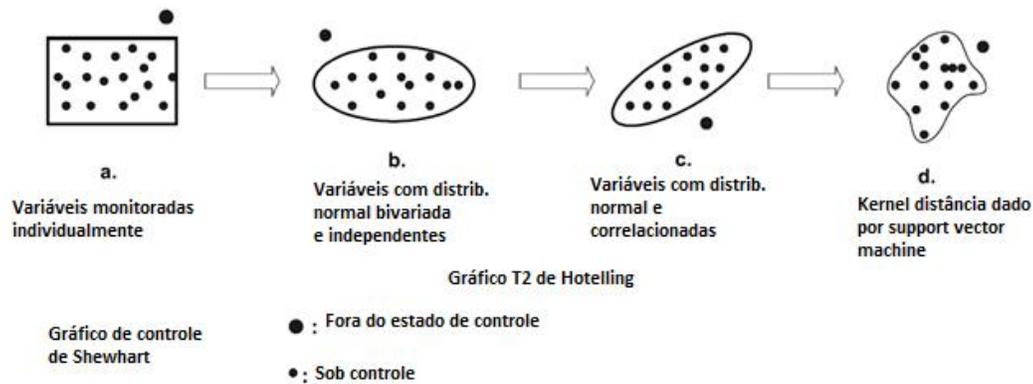


FIGURA 2.28 - Evolução de técnicas de controle estatístico bivariado
FONTE: Adaptado de SUN e TSUNG (2003)

2.2.4 Modelos Híbridos

Os métodos híbridos buscam visualizar uma série temporal composta de uma componente linear autocorrelacionada e uma componente não linear como $y_t = L_t + N_t$, após ajustar uma aproximação \hat{L}_t usando um modelo *ARIMA* se obtém os resíduos $e_t = y_t - \hat{L}_t$, as relações não lineares podem ser modeladas dos resíduos passados, de modo que $e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-k}) + \varepsilon_t = \hat{N}_t + \varepsilon_t$ e se obtém finalmente a previsão $\hat{y}_t = \hat{L}_t + \hat{N}_t$. Alguns resultados têm mostrado que modelos híbridos com Redes Neurais podem produzir uma melhoria em relação aos modelos *ARIMA* tradicionais.

Os modelos *ARIMA* podem representar diferentes tipos de séries temporais, entretanto assumem previamente a forma linear do modelo e que valores futuros da série têm uma relação linear com os valores atuais bem como com o ruído, portanto, podem não ser adequados para complexos problemas não lineares do mundo real. O uso de modelos híbridos ou a combinação de diversos modelos tem se tornado prática comum para melhorar a exatidão de previsões (KHASHEY e BIJARI, 2011).

Zhang (2003) propõe uma metodologia híbrida que combina *ARIMA* e Redes Neurais Artificiais para a previsão de séries temporais. Resultados em alguns conjuntos de dados mostraram uma melhoria na acurácia do desempenho da previsão quando comparado com os dois modelos usados separadamente. A previsão de séries temporais é útil quando há pouco conhecimento disponível sobre o processo

que gera os dados ou quando não há um modelo explicativo satisfatório que relacione as variáveis de previsão com outras variáveis explicativas. Relata ainda que os modelos combinados são mais robustos com relação a uma possível mudança na estrutura dos dados.

Tan *et al.*(2010) propuseram um método de previsão da série de preços no mercado de eletricidade espanhol baseado no uso combinado de um método híbrido que envolve a transformada *Wavelet* e o método ARIMA-GARCH. A série histórica do preço é decomposta em uma série de aproximação e algumas séries de detalhe. Cada subsérie é prevista em separado por um modelo de série temporal adequado. A previsão final é obtida através da composição dos resultados previstos em cada subsérie. O método proposto de previsão é comparado com métodos tradicionais envolvendo Redes Neurais Artificiais e modelagem de séries temporais ARIMA-GARCH. O método híbrido conseguiu capturar melhor a característica complexa de não estacionariedade, não linearidade e volatilidade da série em questão.

Khashey e Bijari (2011) propuseram um modelo híbrido de Redes Neurais Artificiais e modelos ARIMA para produzir previsões mais precisas em séries temporais. Relatam que modelos de Redes Neurais e ARIMA têm obtido sucesso em seus domínios não lineares e lineares, entretanto, nenhum deles é um modelo universal adequado para todas as circunstâncias. Como é difícil possuir o conhecimento completo das características dos dados de um problema real, a metodologia híbrida que tem a capacidade de modelar linear e não linearmente é uma boa estratégia para uso prático. A ideia básica da combinação de modelos para previsão é usar cada característica única do modelo a fim de capturar diferentes padrões nos dados.

Kriechbaumer *et al.*(2014) propuseram um método ARIMA-*Wavelet* para melhoria da previsão de preços de metais. O desempenho da previsão de modelos ARIMA em preços de metais cresceu substancialmente com uma decomposição *Wavelet* prévia da série de preços. Diferentes combinações do tipo de transformada *Wavelet*, função *Wavelet* e número de níveis de decomposição foram testados e isto permitiu aumentar a acurácia (exatidão) da previsão significativamente em relação ao modelo ARIMA clássico.

Liu e Shi (2013) aplicaram modelos ARIMA-GARCH para modelagem e previsão de preços no mercado de eletricidade. Os resultados mostram que tais modelos são uma ferramenta efetiva para modelar e prever a média e a volatilidade dos preços da eletricidade. Relatam que métodos de previsão de séries temporais usam o comportamento passado da série e algumas variáveis exógenas para previsão de valores futuros. Dentre tais métodos destacam-se a inteligência artificial e modelos estatísticos convencionais. Na inteligência artificial as Redes Neurais permitem extrair relações não lineares que governam as entradas e saídas para efetuar previsões. Nos modelos estatísticos cita os modelos ARIMA, ARIMAX e ARIMA-GARCH.

Cao, Qiao e Ren (2009) executaram uma modelagem *Wavelet* seguida de uma rede neural para melhorar a precisão da previsão da deformação de uma barragem. A ideia básica da rede neuro-*Wavelet* é decompor dados de séries com variação no tempo em um conjunto de subséries com o objetivo de tornar a estrutura temporal mais tratável. Após a decomposição *Wavelet*, a rede neural é usada para modelar cada componente e a modelagem global da série se dá pela soma das previsões feitas pela rede em cada componente. Relatam que a dinâmica global da deformação de uma barragem basicamente obedece à propriedade de quase periodicidade devido a variações periódicas aproximadas das principais cargas, por exemplo, pressão da água e temperatura do ar. Os resultados mostraram que nos dados utilizados a metodologia empregada foi mais sensível, efetiva e fornece recursos para a previsão do deslocamento da barragem.

Diversos outros autores têm empregado com sucesso métodos híbridos de previsão de séries temporais. Aplicações na previsão da quantidade de ozônio na superfície em regiões urbanas por Kumar e De Ridder (2010); qualidade do ar (poluentes), por Díaz-Robles *et al.*(2008). No trabalho de Teixeira JR (2013) podem ser encontradas outras referências que fazem uso de combinações de métodos para obter ganhos preditivos na previsão de séries temporais. No contexto do processo de combinação de métodos, os métodos de otimização citados na seção 2.1.11 podem ser fundamentais.

2.2.5 Algumas Experiências em Itaipu e Outras Barragens no Brasil

Segundo Myers e Mariley (1997) com o envelhecimento das barragens, cresceu o interesse na segurança e a necessidade de manutenção. Os responsáveis têm dado maior ênfase na instrumentação automatizada, no monitoramento e em sistemas de informação. Neste trabalho é destacado o planejamento da implementação de um sistema de aquisição automática de dados (ADAS) para monitorar a instrumentação de uma barragem e fornecer alerta precoce de condições de falhas potenciais, a importância do manejo dos dados e da comparação de dados recente com dados históricos para observar tendências.

Itaipu Binacional (2002, p. 5) ao procurar estabelecer critérios para geração de alarmes para controle de instrumentos automatizados, criou três níveis de atenção (controle) baseado em quanto o instrumento ultrapassa um valor pré-estabelecido e do grau de necessidade de averiguação imediata. A mobilização de responsáveis pela supervisão das condições de segurança é feita somente quando, ao menos, dois instrumentos instalados no mesmo bloco e associados entre si apresentam valores dentro das faixas de alerta. Isto é consequência de experiências americanas com automação de instrumentos em barragens, que mostraram que avisos de superação de níveis de alerta passaram a ser muito frequentes, geralmente devido a eventuais falhas do sistema elétrico, causando alarmes falsos e desacreditando o sistema. Ao atingir algum nível de atenção a equipe de segurança de barragem tem uma série de procedimentos a executar. Na análise das causas de ultrapassagem de valores de atenção, qual seja o instrumento a ser avaliado, a primeira tarefa é comparar leituras automatizadas com leituras do relógio comparador convencional, uma vez que a primeira suspeita recai sobre o sistema elétrico de recebimento automático de leituras e o mau funcionamento do instrumento.

No relatório da reunião de consultores civis Itaipu Binacional (2002, p. 15) é informado que as subpressões no contato concreto-rocha, sob a cabeça da montante dos blocos de contraforte e de gravidade aliviada, chegaram a atingir valores acima daqueles estabelecidos pelos critérios de projeto em consequência de uma nítida influência das variações térmicas entre verão e inverno, comprovados por modelos matemáticos de simulação.

Piezômetros instalados na fundação do bloco E-6, mesmo a certa profundidade abaixo do contato concreto-rocha, mostram a influência das variações térmicas. No

QUADRO 2.5 são apresentados os critérios adotados por consultores civis para níveis de atenção (controle) para os piezômetros de Itaipu localizados na fundação de blocos tipo contraforte e gravidade-avaliada (caso do trecho E6).

QUADRO 2.5 - Valores de Atenção para os piezômetros de barragem de gravidade convencional

NÍVEL DE “ATENÇÃO”	CRITÉRIO ADOTADO
A	Critério de projeto para a condição de drenos operantes
B	Média entre os níveis A e C
C	Condição limite de drenos totalmente inoperantes (nível do reservatório)

FONTE: ITAIPU BINACIONAL (2002, p. 15)

Para piezômetros de fundação dos maciços de terra-enrocamento e na interface solo-concreto dos blocos de ligação, Itaipu Binacional (2002, p. 17-18) estabelece os valores de controle em que o nível de atenção A é o nível máximo já observado, o nível C é nível estabelecido pelo critério de projeto e o nível B é a média entre os níveis A e C. Outros instrumentos possuem outros critérios para estabelecer os alertas. A ideia aqui é apresentar alguns trabalhos realizados em Itaipu no intuito de estabelecer valores limite para dados de instrumentação.

Itaipu Binacional (2002) estabeleceu níveis de atenção para: pêndulos diretos e invertidos, extensômetros múltiplos, medidores de junta, piezômetros de fundação, piezômetros da interface de concreto-solo nos blocos de ligação e medidores de vazão. Quando os valores de atenção são ultrapassados deve ser feita, tão logo quanto possível, uma análise de causas num intervalo de tempo que varia de acordo o nível de atenção, conforme apresentado no QUADRO 2.6. A situação de ‘alerta’ é acionada quando ocorre a ultrapassagem dos valores de atenção B, de modo simultâneo, por pelo menos dois instrumentos do bloco chave ou seção instrumentada. A situação de ‘emergência’ é acionada quando ocorre a ultrapassagem dos valores de atenção C, de modo simultâneo, por pelo menos dois instrumentos do bloco chave ou seção instrumentada.

QUADRO 2.6 - Tempo para análise de causas baseado no nível de atenção atingido

NÍVEL DE 'ATENÇÃO'	INTERVALO DE TEMPO
A	7 dias
B	3 dias
C	12 horas

FONTE: ITAIPU BINACIONAL (2002)

Quando a análise de causas não consegue explicar porque valores de atenção foram ultrapassados é sugerida a inspeção detalhada da unidade de aquisição remota (UAR) do bloco ou seção chave em discussão. Um típico painel de alerta é ilustrado na FIGURA 2.29.

Em Itaipu Binacional (1999_c, p. 1) foi desenvolvido um modelo matemático baseado no método dos elementos finitos para estudar o comportamento estrutural do bloco E-6. A calibração do modelo considerou a ação do empuxo da água e comparou resultados de deslocamentos e tensões com as leituras dos instrumentos de auscultação. Para a definição de valores de controle, foram considerados três tipos de condição conforme o QUADRO 2.7.

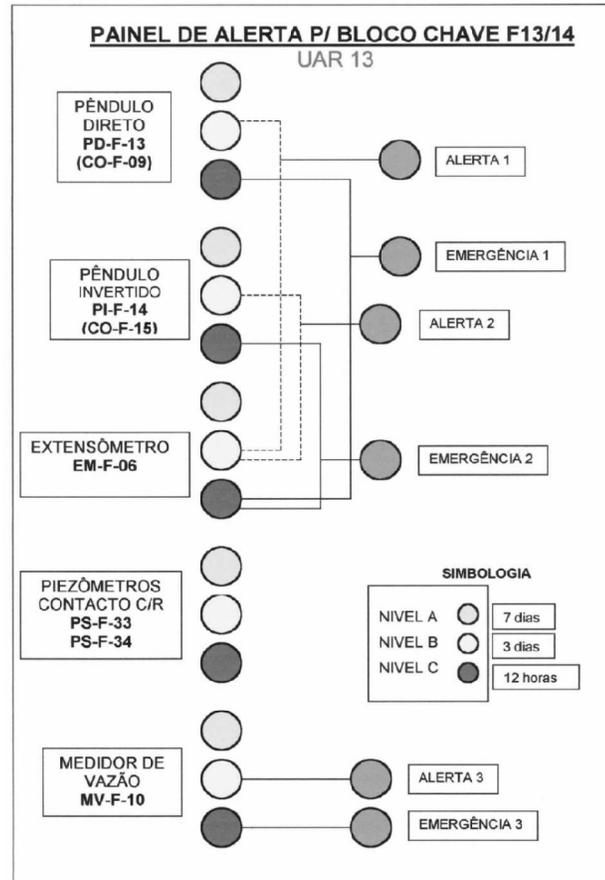


FIGURA 2.29 - Painel de alerta para o bloco Chave F13/14 em Itaipu
Fonte: ITAIPU BINACIONAL (2002, Anexo 3)

QUADRO 2.7 - Condições adotadas no MEF para definir valores de controle no bloco E-6

CONDIÇÃO	CORRESPONDÊNCIA
Carregamento Normal	Variações de temperaturas médias observados no verão e no inverno
Carregamento Excepcional 1	Ação do empuxo da água no nível máximo do reservatório (223m)
Carregamento Excepcional 2	Ação de carga sísmica de 0,05g

FONTE: ITAIPU BINACIONAL (1999_b)

Buzzi (2007) propôs uma metodologia baseada em correlações estatísticas para o entendimento das interações existentes entre as séries temporais de leituras de instrumentos de monitoração geotécnico-estrutural instalados em barragens e analisou a influência que a temperatura ambiente exerce sobre as leituras destes instrumentos para estimar o atraso no tempo de resposta dos instrumentos à

ocorrência de mudanças nos parâmetros ambientais, tais como picos de temperatura. Aplicou seu estudo em leituras de instrumentos instalados nos blocos F 19/20 da barragem de Itaipu. Os resultados permitiram quantificar correlações e intervalos de tempo relevantes à segurança da barragem.

Villwock (2009) com o objetivo de melhor “explicar” o comportamento da barragem da Usina Hidrelétrica de Itaipu apresentou uma metodologia, enquadrada na área de Descoberta de Conhecimento em Bases de Dados, com o intuito de realizar a hierarquização de instrumentos de monitoramento de barragens, maximizando a eficácia e eficiência das análises das leituras, através da identificação de grupos de instrumentos semelhantes e, também, detectando os principais instrumentos. Na mineração dos dados, a tarefa foi o agrupamento de padrões com estatística multivariada (Ligação Simples, Média, Completa e Método Ward), Redes Neurais de Kohonen e da área de Metaheurísticas foi proposto um método de Agrupamento baseado no Algoritmo de Colônias de Formigas. Na aplicação da Análise de Agrupamento aos dados de instrumentação de Itaipu, mostrou-se que é possível encontrar justificativas técnicas para a formação dos grupos, inclusive identificando um grupo de hastes de maior importância. A técnica se mostrou eficaz para realizar a hierarquização das hastes de extensômetros e de identificar o grupo de hastes de maior importância.

Rosso *et al.* (1995, p. 2-3) avaliaram a influência térmica sobre os deslocamentos da crista da barragem, subpressões na região de contato concreto-rocha e infiltrações através do concreto e da fundação. Apontam que as barragens (como Itaipu) do tipo gravidade aliviada são muito sensíveis a variações térmicas. A retroanálise de dados de instrumentação colabora no sentido de entender o mecanismo de atuação da influência térmica nos deslocamentos da barragem, subpressões na região de contato concreto-rocha, sob a cabeça da montante.

Rosso *et al.* (1995, p. 5) concluíram que a face de montante sofre menos influência da variação da temperatura, quanto a deslocamentos horizontais da crista da barragem de concreto, por estar em contato com a água do reservatório. Já a face da jusante apresenta variações térmicas mais significativas e isto explica o deslocamento horizontal da crista da barragem para jusante no inverno e para

montante no verão. Ressaltam ainda que alguns modelos teóricos não consideraram o efeito térmico sazonal.

Diante da constatação de subpressões superiores ao que se previa em projeto, Rosso *et al.* (1995, p. 8) decidiram que as subpressões previstas em projeto deveriam passar a ser entendidas como valores de controle e não mais como valores limites. A conclusão para o aparecimento de subpressões elevadas na região de contato concreto-rocha, sob a cabeça de montante, é a influência térmica no inverno de distensões no concreto. Neste caso, nota-se que há a intenção de estabelecer, novamente, níveis de 'atenção' ou de controle que diferem dos critérios iniciais de projeto.

Vale ressaltar que a magnitude de variação da subpressão devido à sazonalidade passou a integrar o critério de projeto para novas barragens similares, muito embora, especialistas apontam que em termos de obras de barragens, cada caso deve ser analisado dentro de suas especificidades.

De modo geral, Rosso *et al.*(1995, p. 11) constataram que as estruturas de contraforte e gravidade aliviada em Itaipu, são influenciadas pelas variações térmicas ambientais, causando deslocamentos da crista da barragem, tensões no concreto, infiltrações e subpressões no contato concreto-rocha, sob a cabeça da montante. Em particular, altas subpressões indicadas pelos piezômetros provocaram a revisão dos critérios de projeto para subpressões em determinadas regiões. Os blocos de contrafortes (como no trecho E6) são os mais expostos às influências térmicas ambientais.

As leituras da maioria dos instrumentos instalados em usinas hidrelétricas da Companhia Energética de São Paulo não possuíam valores limites associadas as suas leituras. Kuperman *et al.* (2004) e Kuperman *et al.*(2007) apresentaram um critério desenvolvido para estabelecer valores limite para as leituras de instrumentos em barragens de usinas com base em intervalos de confiança, correlações e regressões múltiplas e médias móveis. Com isso o processo de monitoramento tornou-se mais confiável e ajudou a alterar a frequência de leituras e a redução de custos. O trabalho utilizou-se da hipótese de que os dados seguiam uma distribuição normal de probabilidades. Relatam ainda que alguns instrumentos do mesmo tipo localizados em diferentes locais da barragem foram comparados e que suas variações

estão associadas com a oscilação do nível de água do reservatório ou com a temperatura ambiente. Alguns valores limite, chamados de valores de referência (inferior, médio e superior) no trabalho, foram estabelecidos com base na média e no desvio padrão das leituras e construção de intervalos de confiança de 95% para a distribuição normal ou *t*. Os valores de referência (ou limites) servem para a indicação de erros de medição e ocorrência de anormalidades no instrumento ou da estrutura.

O objetivo da instrumentação em barragens é medir as cargas que afetam a estrutura e os seus efeitos. O monitoramento dos instrumentos ao longo dos anos permite a detecção precoce de problemas em potencial e a comparação do comportamento das estruturas com o que foi projetado. Instrumentos convenientemente instalados com leituras adequadas são fundamentais para estabelecer o desempenho da barragem. Por outro lado, a instrumentação instalada em uma barragem não é garantia da detecção de problemas (KUPERMAN *et al.*, 2004); KUPERMAN *et al.*, 2007).

Não é possível aplicar valores limites estatísticos obtidos em outras barragens uma vez que cada barragem possui um comportamento único que reflete condições específicas do local e do projeto. Podem ser adotados três modelos para lidar com a questão do estabelecimento de valores limites. O modelo determinístico se utiliza de informações completas do projeto, das cargas reais, propriedades dos materiais e da fundação, caso disponíveis. O modelo estatístico se utiliza de séries históricas de leituras, cargas e seus efeitos e métodos híbridos, quando existem dados para estimar algumas cargas e seus efeitos, mas não é possível estimar outras cargas atuantes (KUPERMAN *et al.*, 2004).

Para se estabelecer critérios estatísticos de valores de controle para instrumentos, uma hipótese básica é que se os valores medidos pelos mesmos instrumentos continuam a variar dentro de certo domínio ao longo dos anos, mantido as demais condições similares ao passado, o comportamento estrutural se manterá dentro de condições 'normais' (KUPERMAN *et al.*, 2004).

3 MATERIAL E MÉTODOS

Neste capítulo são apresentadas as justificativas das escolhas das técnicas e os materiais utilizados neste trabalho (seção 3.1), que incluem os dados e os *softwares* utilizados para aplicação dos métodos descritos no capítulo 2. Considerando que houve trabalhos univariados e multivariados, as subseções 3.2.1 e 3.2.2 fazem um detalhamento da metodologia empregada separadamente para cada trabalho.

Este trabalho limitou-se ao monitoramento do gráfico de controle da média ou do vetor de médias de um processo, muito embora existam gráficos para o controle da variância, da amplitude, entre outros.

Além disso, no caso univariado também se limitará aos gráficos de Shewhart, em detrimento de outros gráficos de controle univariados como os de somas cumulativas (CUSUM) e de médias móveis exponencialmente ponderadas (EWMA) mais indicados quando se deseja detectar mudanças no processo da ordem de $1,5\sigma$ ou menos, tendo em vista que tais procedimentos são mais indicados para detectar erro tipo II.

Para o tratamento paramétrico escolheu-se a modelagem ARIMA(X)-GARCH por ser consagrada na modelagem do comportamento linear de séries temporais e por permitir tratar com séries temporais eventualmente não estacionárias, que apresentem estrutura de autocorrelação, com sazonalidades e variância não constante e possibilidade de lidar com variáveis exógenas.

As Redes Neurais Artificiais foram escolhidas por seu sucesso na modelagem de comportamento não linear de séries temporais e de não requerer hipóteses restritivas sobre a série temporal. Diante de diversas formas de arquitetura, foram empregadas as Redes Neurais *feedforward* com uma camada escondida, considerando o teorema de aproximação universal de Cybenko, conforme Teixeira JR. (2013), que diz que uma rede *feedforward* com uma única camada escondida contendo um número finito de neurônios, pode aproximar funções contínuas em subconjuntos compactos de \mathbb{R}^n , segundo hipóteses básicas sobre a função de ativação. O teorema indica que geralmente uma camada escondida é suficiente para obter aproximações tão boas quanto se queira.

As combinações lineares híbridas de métodos de previsão de séries temporais tem mostrado eficiência no sentido de produzir melhores previsões e agregar informações de diferentes fontes e, portanto, foram selecionadas neste trabalho.

As *Wavelets* foram usadas, pois a decomposição de séries temporais permite sua análise em diferentes escalas de frequência e de tempo e isto é capaz de produzir previsões mais precisas.

Para o tratamento não paramétrico dos dados foram escolhidos os métodos *Kernel Density Estimation (KDE)* e *Bootstrap* por serem técnicas robustas, isto é, capazes de gerar limites de controle “próximos” dos valores exatos quando se lida com dados de uma distribuição específica e que podem ser aplicados em situações em que não se conhece a distribuição subjacente dos dados.

O trabalho teve seu foco principal na análise dos erros do tipo I, ou falsos alarmes uma vez que o objetivo primordial é estabelecer os limites de controle, considerando o tempo de estudo até agora, que é o caso específico dos dados disponíveis. Isto é fundamental para um posterior estudo de erros tipo II, ou falsos verdadeiros.

No caso de gráficos multivariados o trabalho se restringe ao gráfico T^2 de Hotelling por se tratar do procedimento mais familiar de controle multivariado. Procedimentos análogos aos de somas cumulativas e de médias móveis exponencialmente ponderadas não foram tratados neste trabalho.

Gráficos multivariados das componentes principais são utilizados por permitirem monitorar processos onde a variabilidade não é distribuída uniformemente entre todas as variáveis, onde haja correlação e por permitirem uma ‘redução’ nos dados, a descoberta de novas associações e uma visão mais ampla do monitoramento das variáveis. Formas alternativas de abordagem multivariada como análise de componentes principais não lineares e vetores autorregressivos (VAR) ficaram de fora do escopo deste trabalho.

3.1 MATERIAL

Foram usados dados provenientes da leitura de instrumentos de monitoramento da barragem da usina hidrelétrica de Itaipu Binacional, localizada no Rio Paraná, entre o Brasil e o Paraguai. Itaipu é a maior geradora de energia elétrica do mundo (com exceção de 2014), sua geração atende cerca de 17% da

energia elétrica anualmente consumida no Brasil e 75% do Paraguai.

De acordo com Rosso *et al.*(1995, p. 2-3) a barragem principal de Itaipu é constituída por blocos tipo gravidade aliviada com 34 m de largura e 612 m de extensão total e na região do leito do Rio Paraná atinge a altura máxima de 196 m. Ao lado desta estrutura, a barragem principal de Itaipu é integrada por 83 blocos tipo contraforte, com 17 m de largura, altura média de 60 m e extensão total de 1411 m. Ambas as estruturas possuem instrumentação detalhada e acompanhamento pormenorizado do seu comportamento durante o período operacional da barragem com vistas a verificar as hipóteses de projeto e supervisionar sua segurança. A FIGURA 3.1 exibe um arranjo geral das estruturas de concreto na barragem de Itaipu.

Devido à grande extensão da barragem de Itaipu, por sugestão da equipe de segurança, como o trecho E é constituído de 6 blocos e o bloco E-6 é considerado um bloco chave (mais instrumentado), este trabalho limitou-se a este bloco. O bloco E-6 está localizado na ligação direita da barragem principal, na região indicada na FIGURA 3.2. O bloco E-6 é um dos blocos que compõe a ligação entre os blocos de contraforte e a barragem principal e possui alta densidade de instrumentação. Sua altura a montante vai da elevação 140 m até 225 m e à jusante da elevação 125 m até 225 m, portanto, possui 85 m de altura a montante e 100 m de altura a jusante e a largura da 'cabeça' do bloco é de 17 m. Uma parte do local onde são realizadas as leituras da instrumentação do bloco E-6 pode ser vista na FIGURA 3.3.

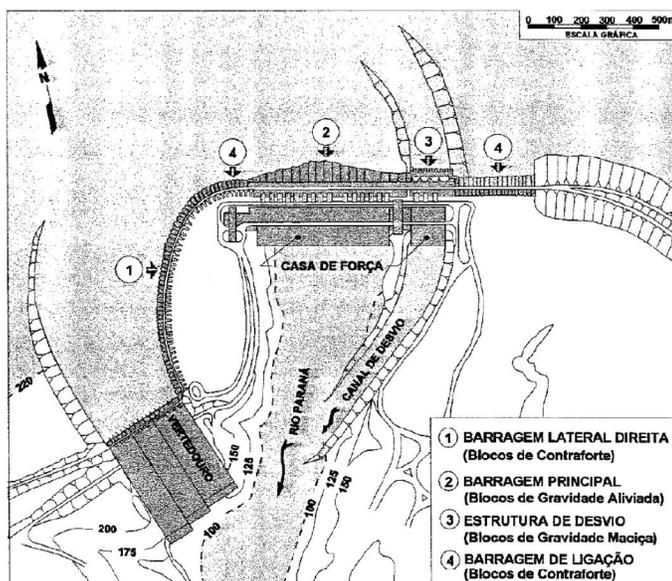


FIGURA 3.1 - Arranjo geral do projeto de Itaipu
FONTE: ROSSO *et al.* (1995)



FIGURA 3.2 - Barragem da Usina de Itaipu e o Trecho E em destaque
FONTE: Itaipu Binacional



FIGURA 3.3 - Parte da instrumentação do bloco E-6
FONTE: O autor

No bloco E-6 estão disponíveis dados de mais de 40 instrumentos com 11 finalidades. Os instrumentos estão descritos no QUADRO 3.1 e na FIGURA 3.4 pode ser visto o bloco e toda sua instrumentação.

Considerando o tempo disponível, prezando pela qualidade da interpretação e a experiência da equipe de Itaipu, resolveu-se trabalhar somente com os seguintes instrumentos: piezômetros, pêndulos, extensômetros e bases de alongômetros, além de instrumentos que medem grandezas relacionadas a temperaturas (interna, externa e da água) e nível do reservatório.

QUADRO 3.1 - Instrumentos localizados no bloco/trecho E6

	INSTRUMENTO	QUANTIDADE
1	Piezômetro <i>Standpipe</i>	9
2	Extensômetro Múltiplo	4
3	Base de alongâmetro	3 (1x3)
4	Medidores triortogonais	2 (x,y,z)
5	Pêndulo Direto	1
6	Roseta de Tensômetro	1
7	Termômetro Interno	3
8	Tensômetro de Concreto	5
9	Termômetro superficial	1
10	Medidor de vazão	1
11	Medidor de junta	1

FONTE: Equipe de segurança de barragem de Itaipu

Também foi utilizada uma série de vazões médias mensais naturais no Rio Paraná, posto 266, próximo a Itaipu Binacional, disponibilizado pelo ONS (OPERADOR NACIONAL DO SISTEMA ELÉTRICO, 2016) com vistas a possibilitar sua utilização em modelagem matemática, pesquisa e desenvolvimento, devido a importância da vazão afluyente disponível para a produção de energia em Itaipu.

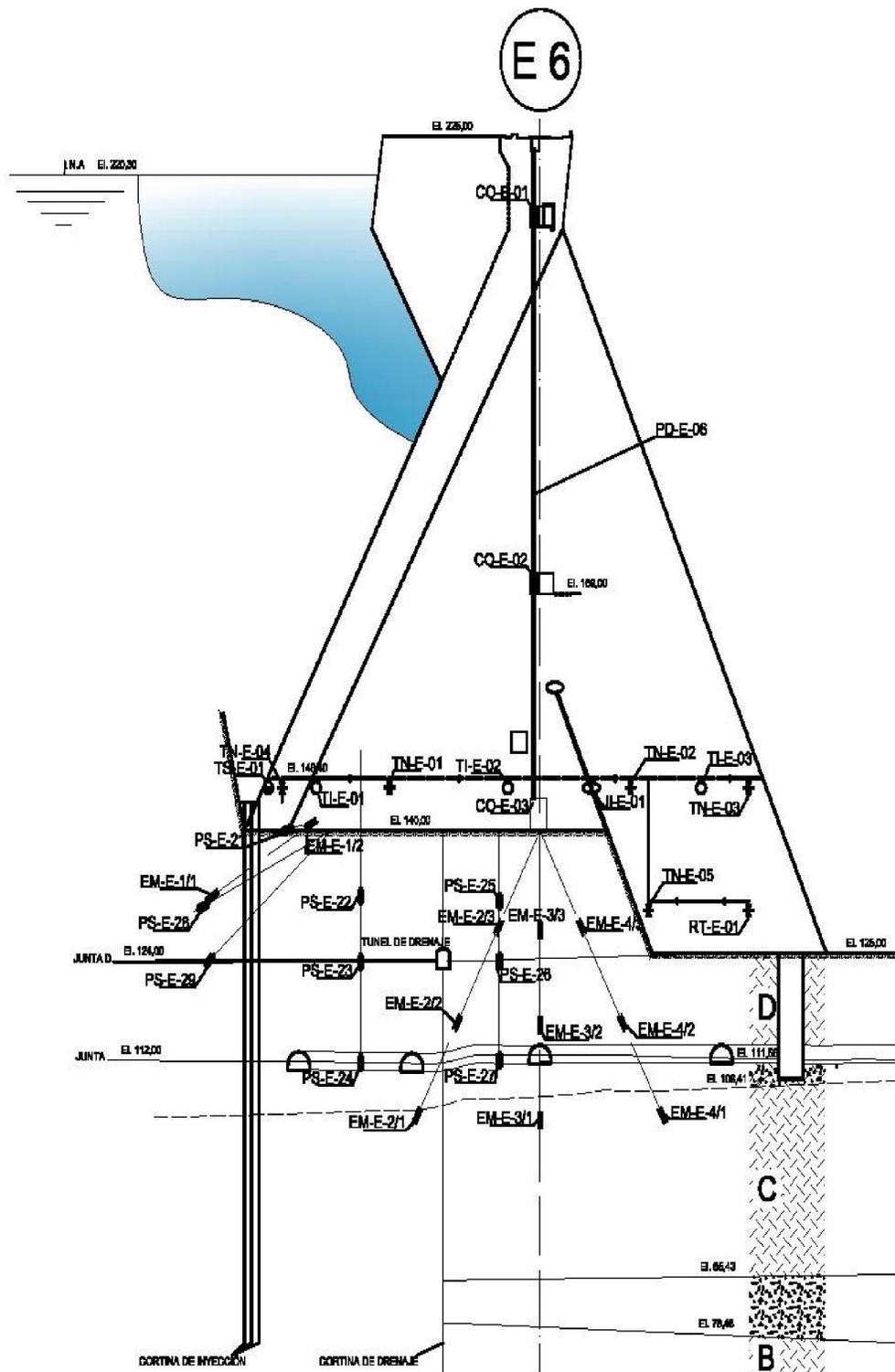


FIGURA 3.4 - Instrumentação do Bloco E-6 em Itaipu
 FONTE: Equipe de segurança de barragem de Itaipu

Um piezômetro é um dispositivo de medição que é vedado e enterrado de modo que responda apenas à pressão da água subterrânea em torno de si, somente na elevação em que estiver enterrado. A elevação piezométrica ou nível piezométrico é a elevação da água em um piezômetro. O nível piezométrico e as pressões são raramente constantes ao longo do tempo. Forças naturais tais como precipitação, evaporação, pressão atmosférica e infiltração podem causar grande variação no nível do lençol freático (ou das águas subterrâneas). Normalmente um piezômetro, quando há mudança na pressão da água subterrânea, possui uma pressão hidrostática defasada no tempo, que é o tempo que ele responde a esta mudança. Isto depende da permeabilidade do solo, tipo de piezômetro, dentre outros aspectos (USACE, 1995). A FIGURA 3.5 mostra a resposta defasada de um instrumento.

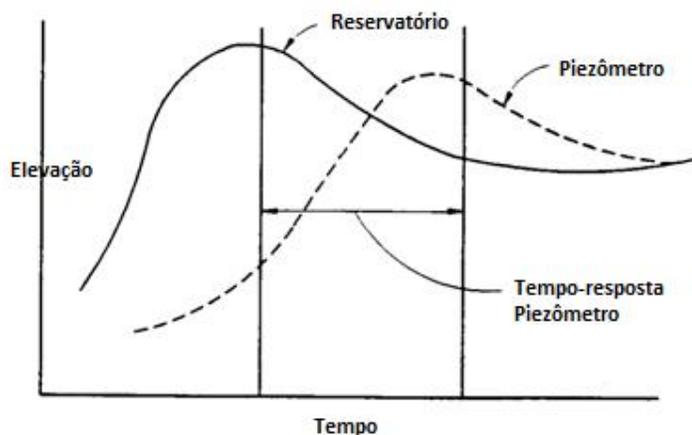


FIGURA 3.5 - Resposta defasada de um piezômetro em relação ao nível do reservatório
FONTE: Adaptado de USACE (1995)

A FIGURA 3.6 exibe o local onde são feitas as leituras dos piezômetros PS-E-28 e PS-E-29 no trecho E da barragem da usina de Itaipu.



FIGURA 3.6 - Local de leituras dos piezômetros 28 e 29 no trecho E
FONTE: O Autor

Embora estejam disponíveis dados históricos destes instrumentos desde o período de enchimento do reservatório de Itaipu em 1984, devido a diferentes períodos de leitura pelos quais a instrumentação já passou (periodicidade menor no início do período operacional), dependendo da etapa do trabalho realizado, resolveu-se trabalhar com os dados a partir de determinado período em diante, para padronizar a periodicidade.

Uma parte dos dados de Itaipu, cerca de 10%, tem leitura automatizada e os dados usados neste trabalho foram oriundos de leituras automáticas ou manuais, conforme o que foi fornecido. Os instrumentos com os quais se trabalhou, foram: piezômetros, pêndulo direto, extensômetros e bases de alongâmetro.

A metodologia deve descrever a seleção e a preparação dos dados, o que inclui quais variáveis devem ser selecionadas e o pré-tratamento dos dados como centralização, ajuste de escala e filtragem, isto é vital para o sucesso de uma aplicação (MILETIC *et al.*, 2004, p. 825). Neste trabalho se buscou um equilíbrio na escolha e o uso do número de variáveis. Muitas variáveis podem 'diluir' a informação e dificultar a identificação de informações úteis, por outro lado, poucas variáveis podem prejudicar ou levar a interpretações da solução incorretas pelo fato de excluir variáveis úteis.

Neste trabalho foram utilizados diversos *softwares* com diversas finalidades. A seguir segue uma breve descrição do *software* e da sua finalidade.

- o *STATGRAPHICS Centurion XVI*®, de acordo com Statgraphics Centurion XVI (2014), é um *software* de para análise estatística, previsão de séries temporais, controle de qualidade e possui mais de 180 procedimentos estatísticos, dentre eles, métodos multivariados e técnicas não paramétricas. O programa foi utilizado, sobretudo, para construção de gráficos de controle.
- O MATLAB® (versão R2013a, 8.1.0.604) é um ambiente interativo direcionado ao cálculo numérico com matrizes amplamente utilizado por cientistas de várias áreas. Ele permite a resolução de problemas numéricos complexos criando-se códigos em sua própria linguagem de programação (MATHWORKS, 2015). Neste trabalho seu uso teve a finalidade de obter a decomposição *Wavelet* e adaptar Redes Neurais as séries temporais, além de integração numérica.
- O programa *Econometric Views* (EVIEWES® 8) é um pacote estatístico desenvolvido para manipular dados de séries temporais e, apesar de desenvolvido para aplicações em economia, pode ser utilizado em diversas áreas (PINTO e SILVA, 2015). Sua principal finalidade foi de estimar modelos ARIMA, ARIMAX-GARCH e fazer previsões de séries temporais.
- As planilhas eletrônicas *Excel*® foram utilizadas para armazenamento de dados das séries temporais para que se pudessem exportar dados para outros *softwares*, bem como, receber resultados obtidos de outros *softwares*.

3.2 MÉTODOS

Como este trabalho engloba tanto estratégias univariadas como estratégias multivariadas de tratamento das séries temporais de dados de leituras de instrumentos de monitoramento do trecho E6 da barragem da Usina Hidrelétrica de Itaipu e de uma série de vazões de afluentes do Rio Paraná do Operador Nacional do Sistema, os métodos empregados em cada caso (univariado/multivariado) serão descritos separadamente.

3.2.1 Caso Univariado

3.2.1.1 O caso do pêndulo direto

Neste trabalho foi realizado um estudo comparativo entre diversas abordagens para previsão estática da série temporal de dados da leitura de um pêndulo direto, que

mede o deslocamento da crista da barragem na direção do fluxo do rio, da barragem da usina hidrelétrica de Itaipu, localizada no Rio Paraná, entre o Brasil e o Paraguai.

O objetivo foi comparar o desempenho das previsões realizadas em diversas abordagens e mostrar que o uso da decomposição *Wavelet* e da combinação de métodos preditivos como *ARIMA-GARCH* e Redes Neurais Artificiais pode acarretar melhorias relevantes no desempenho da previsão do comportamento de barragens e auxiliar, conseqüentemente, no fornecimento de alertas prévios de eventuais anormalidades.

Em grandes barragens, há instrumentos instalados nas estruturas de concreto e nas fundações da barragem. Em especial, o pêndulo direto é um instrumento instalado na estrutura de concreto e é responsável por medir o deslocamento da crista da barragem na direção do fluxo e normal a ele em relação a sua base.

As subpressões e infiltrações nas barragens de concreto não podem ser previstas com acurácia por modelos numéricos, como Métodos de Elementos Finitos (MEF) que se baseiam nas leis físicas que governam o fenômeno envolvido. Por isso, a disponibilidade de dados de monitoramento tem difundido a aplicação de modelos estatísticos para a previsão do comportamento da barragem e estes têm servido como um complemento às inspeções visuais e aos modelos numéricos no suporte a tomada de decisões (SALAZAR *et al.*, 2015). Portanto, como, em geral, as leituras de instrumentos de monitoramento de barragens são avaliadas com valores definidos em projeto, como MEF, o seu domínio de variação pode não ser mais realístico durante seu período operacional e são necessárias outras formas de monitoramento.

Os dados fornecidos por Itaipu são compostos por 280 observações mensais do pêndulo direto localizado no trecho E da barragem, registradas no período de julho de 1991 a outubro de 2014, ou seja, um vetor com 280 dados. A definição deste período foi em função da padronização da periodicidade. A amostra foi separada em amostra de treinamento, com 240 dados do período de julho de 1991 a junho de 2011, usada para construção e estimativa dos parâmetros dos modelos e amostra de validação, com 40 dados do período de julho de 2011 a outubro de 2014, para avaliar o desempenho em termos da capacidade preditiva das diversas abordagens (modelos) propostas. Foram empregadas quatro medidas de avaliação de

desempenho: MAPE (Erro Absoluto Percentual Médio), MSE (Erro Quadrático Médio), MAE (Erro Absoluto Médio) e R^2 (Coeficiente de Explicação).

Com a intenção de comparar os modelos, as técnicas testadas utilizando *Wavelet*, *ARIMA-GARCH* e RNA's, bem como as combinações destas, foram adotadas onze abordagens conforme o QUADRO 3.2. Os modelos *ARIMA* e *ARIMA-GARCH* foram determinados através do *software EViews*, versão 8, e a escolha do modelo final se deu em termos da ordem de diferenciação para obter estacionariedade, sazonalidade, significância dos termos autorregressivos e de médias móveis, diagnóstico dos resíduos ordinários padronizados e resíduos quadráticos (função de autocorrelação e autocorrelação parcial), e da estatística de Durbin-Watson.

Para os modelos de Redes Neurais foram usadas RNA's do tipo *feedforward*, desenvolvidas no *software Matlab*. Os parâmetros selecionados foram: janela de previsão variando de 1 a 13 e número de neurônios na camada escondida variando de 1 a 10. Amostra de treinamento ($200/280 = 0,71$), amostra de validação ($40/280 = 0,145$) e amostra de teste ($40/280 = 0,145$). A medida do desempenho para a seleção da rede ótima foi o *MSE*.

No caso da decomposição *Wavelet* da série original foi considerado somente um tipo de decomposição, em que, por simplicidade, se utilizou o nível de decomposição $p = 2$ e a base de *Haar*, tanto para a posterior modelagem *ARIMA-GARCH* quanto para modelagem com rede neural, da qual se obteve uma série de aproximação e duas séries de detalhes. A decomposição foi feita no *MatLab*, v.8.

QUADRO 3.2 - Abordagens testadas

ABORDAGEM	DESCRIÇÃO
1	Modelagem <i>ARIMA</i> da série original.
2	Modelagem com Rede Neural da série original.
3	Modelo híbrido de combinação linear <i>ARIMA</i> + Rede Neural (em paralelo) da série original.
4	Modelo híbrido de combinação linear <i>ARIMA</i> (série original) + Rede Neural (sequencial – rede aplicada nos resíduos <i>ARIMA</i> da série original).
5	Decomposição <i>Wavelet</i> (série original) + Modelagem <i>ARIMA</i> – <i>GARCH</i> em cada componente.
6	Decomposição <i>Wavelet</i> (série original)+ Modelagem com rede neural em cada componente.
7	Decomposição <i>Wavelet</i> (série original) + Modelo híbrido em paralelo da combinação linear ótima da soma da combinação ótima <i>ARIMA</i> – <i>GARCH</i> de todas as componentes com a combinação ótima de REDE NEURAL de todas as componentes.
8	Decomposição <i>Wavelet</i> (série original) + Modelo híbrido em paralelo da soma das combinações ótimas (<i>ARIMA</i> – <i>GARCH</i> + <i>RNA</i>) em cada componente.
9	Decomposição <i>Wavelet</i> (série original) + Modelo híbrido em paralelo da combinação linear ótima da soma das combinações ótimas (<i>ARIMA</i> – <i>GARCH</i> + <i>RNA</i>) em cada componente.
10	Decomposição <i>Wavelet</i> (série original) + Modelagem <i>ARIMA</i> – <i>GARCH</i> em cada componente seguida da modelagem dos resíduos de cada componente com Redes Neurais.
11	Modelo híbrido da combinação da soma das previsões <i>ARIMA</i> – <i>GARCH</i> das componentes com as previsões de Redes Neurais dos resíduos em cada componente.

Fonte: O autor

A FIGURA 3.7 apresenta um fluxograma da abordagem 7, mas serve de ilustração para representar a maioria das abordagens, algumas das quais são parte da abordagem 7 ou bastante semelhantes a esta. Para as combinações lineares foi feita uma regressão da variável dependente sobre as variáveis independentes. Por exemplo, a combinação ótima *ARIMA-GARCH* da abordagem 7, chamada de Combinação Linear *Arima-Garch* (*CL AG*) na FIGURA 3.7, usa a série original como variável dependente e a modelagem com *Arima-Garch* de cada componente (A_1, D_2, D_3) da decomposição *Wavelet* feita, denotadas por $AG A_1, AG D_2, AG D_3$, como variáveis independentes.

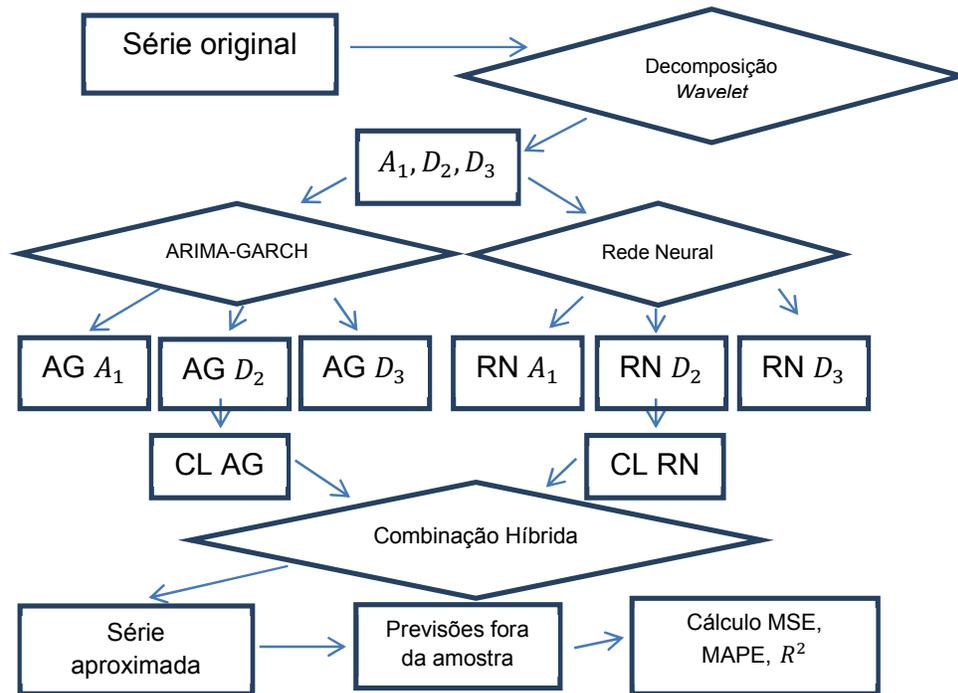


FIGURA 3.7 - Fluxograma para a abordagem 7 do QUADRO 3.2

Fonte: O autor

Os resultados obtidos com essa metodologia univariada estão apresentados na seção 4.1.1 (página 162).

3.2.1.2 O caso da série da vazão de afluentes no Rio Paraná

Neste trabalho foi proposta uma metodologia híbrida para combinar previsões de séries temporais estocásticas que primeiramente realiza uma decomposição *Wavelet* de nível p da série temporal, produzindo $p+1$ componentes *Wavelet* que são modeladas individualmente por meio de um modelo de Box e Jenkins e uma Rede Neural Artificial, a fim de capturar, respectivamente, estruturas lineares e não lineares de autodependência para, então, serem linearmente combinadas, fornecendo previsões híbridas para cada componente. Finalmente, todos estes são combinados linearmente. Para a avaliação, foram utilizados os modelos de Box e Jenkins (BJ), Redes Neurais Artificiais (RNA), e sua combinação tradicional Linear (CL1) de previsões; Redes Neurais integradas com a decomposição *Wavelet* (RNA-

WAVELET), modelo de BJ integrado com a decomposição *Wavelet* (BJ-*WAVELET*), e sua combinação linear convencional (CL2) de previsões. Todos os métodos de previsão foram aplicados à série temporal mensal da vazão de afluentes do Rio Paraná num posto de medição localizado próximo a usina de Itaipu, cujos dados foram obtidos no *site* do Operador Nacional do Sistema Elétrico (2016) e a metodologia híbrida proposta teve desempenho preditivo mais elevado do que os outros.

Para ilustrar a metodologia híbrida proposta foi modelada a série temporal da vazão média de afluentes no Rio Paraná, num posto de medições próximo a usina de Itaipu, no período de janeiro de 1970 até dezembro de 2010 (ou seja, um vetor com 492 observações). Sua escolha foi devido a sua relevância e a dificuldade de modelagem. Para fins experimentais, 80% dos dados iniciais foram utilizados na formação de modelos individuais (Box & Jenkins e Redes Neurais Artificiais), bem como para a obtenção dos parâmetros ótimos das combinações de previsões; 10% dos dados foram utilizados na subsequente amostra de validação; e 10% do restante na amostra de teste. Foram feitas previsões um passo em frente, em um horizonte de 49 passos à frente de previsão. Para avaliar o desempenho dos métodos preditivos foram usadas as estatísticas MAE (erro médio absoluto) e MAPE (erro médio percentual absoluto) dos resíduos.

O método proposto pode ser descrito nos seguintes passos básicos:

- (1) Obter os valores da série temporal a ser modelada;
- (2) Realizar a decomposição *Wavelet* de nível p , gerando uma componente de aproximação e p componentes de detalhe;
- (3) Modelar cada uma das $p + 1$ componentes *Wavelet*, (CW) do passo (2) através de um modelo de Box e Jenkins e de uma rede neural artificial;
- (4) combinar as previsões para cada CW originárias de ambos os previsores individuais mencionados (BJ e RNA) por meio de previsões lineares tradicionais, onde o ajuste dos parâmetros é feita por meio de programação não linear, uma vez que a função objetivo será minimizar o erro quadrático médio de previsão;
- (5) combinar linearmente as previsões de cada CW híbrida geradas no passo (4) gerando a previsão desejada para a série temporal. Os resultados desta metodologia univariada estão apresentados na seção 4.2 (página 168).

3.2.2 Caso multivariado

3.2.2.1 Estatísticas T^2 e Q no estudo dos piezômetros

O primeiro trabalho multivariado foi desenvolvido com dados reais provenientes da instrumentação do trecho E da usina hidrelétrica de Itaipu. Para este trabalho foram escolhidos sete piezômetros em virtude da quantidade e da importância da sua função em medir subpressões na barragem, dois piezômetros do total de nove existentes não foram usados pois um deles é considerado inoperante por apresentar leituras constantes e o outro por ter tido um grande período de leituras ausentes.

O objetivo foi propor um modelo estatístico multivariado para o monitoramento de instrumentos de auscultação de barragens via gráficos de controle e análise de componentes principais que busca separar o efeito de variáveis ambientais sobre a leitura dos instrumentos de outras fontes de variabilidade pelo uso das estatísticas T^2 e Q e estabelecer valores de controle para o monitoramento de futuras observações.

Como a periodicidade das leituras desde o período do enchimento do reservatório em 1984 variou, foi escolhido para análise o período de 2001 até 2013 quando as leituras foram aproximadamente quinzenais para todos os piezômetros. Este período gerou 319 leituras para cada instrumento, daqui por diante denominados de piezômetros P_1, P_2, \dots, P_7 . Para a fase I, de teste de ajustamento (retrospectiva) do modelo foram selecionadas 300 leituras e para a fase II de validação (perspectiva) do modelo foram selecionadas as 19 leituras restantes. É importante citar que nesta fase I os dados são considerados sob controle do ponto de vista estatístico. Portanto, a matriz de dados consiste de 319 linhas e 7 colunas.

As etapas do método consistem em:

- Coleta, tratamento e padronização dos dados;
- Testar a normalidade multivariada;
- Escolher a probabilidade de erro tipo I e a estimativa da matriz de covariância;
- Construir gráficos de Shewhart para cada variável individualmente;
- Construir o gráfico T^2 para o conjunto total de variáveis;
- Extração e seleção do número de componentes principais e teste de normalidade;

- Construção do gráfico T^2 e da elipse de controle das componentes principais nas fases I e II;
- Cálculo dos resíduos e da estatística Q ;
- Construção do gráfico de Q ;
- Interpretação dos resultados.

Os resultados desta metodologia multivariada estão apresentados na seção 4.2.1 (página 168).

3.2.2.2 O caso dos piezômetros do trecho E-6 via KDE

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, várias abordagens têm sido propostas. Cheng e Zheng (2013) criaram 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 (PHALADIGANON *et al.*, 2013). Simulações da estatística T^2 de Hotelling mostraram que o gráfico proposto tem melhor desempenho na ausência da normalidade.

Neste trabalho, o objetivo foi estabelecer limites de controle para o gráfico multivariado dos resíduos de componentes principais (ACP) retidas de um processo multivariado que consistia dos dados de leituras de sete piezômetros localizados no trecho E da barragem da usina de Itaipu. 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 . Foi realizada uma comparação com os limites obtidos com a hipótese de normalidade multivariada (MVN) para dois tipos de estimativas da matriz de covariância dos dados.

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, \dots, p_7) localizados no trecho E6 no período de 2001 a 2013, por ser mais atual e de periodicidade aproximadamente quinzenal para todos os instrumentos integrantes da análise. 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 3.8.

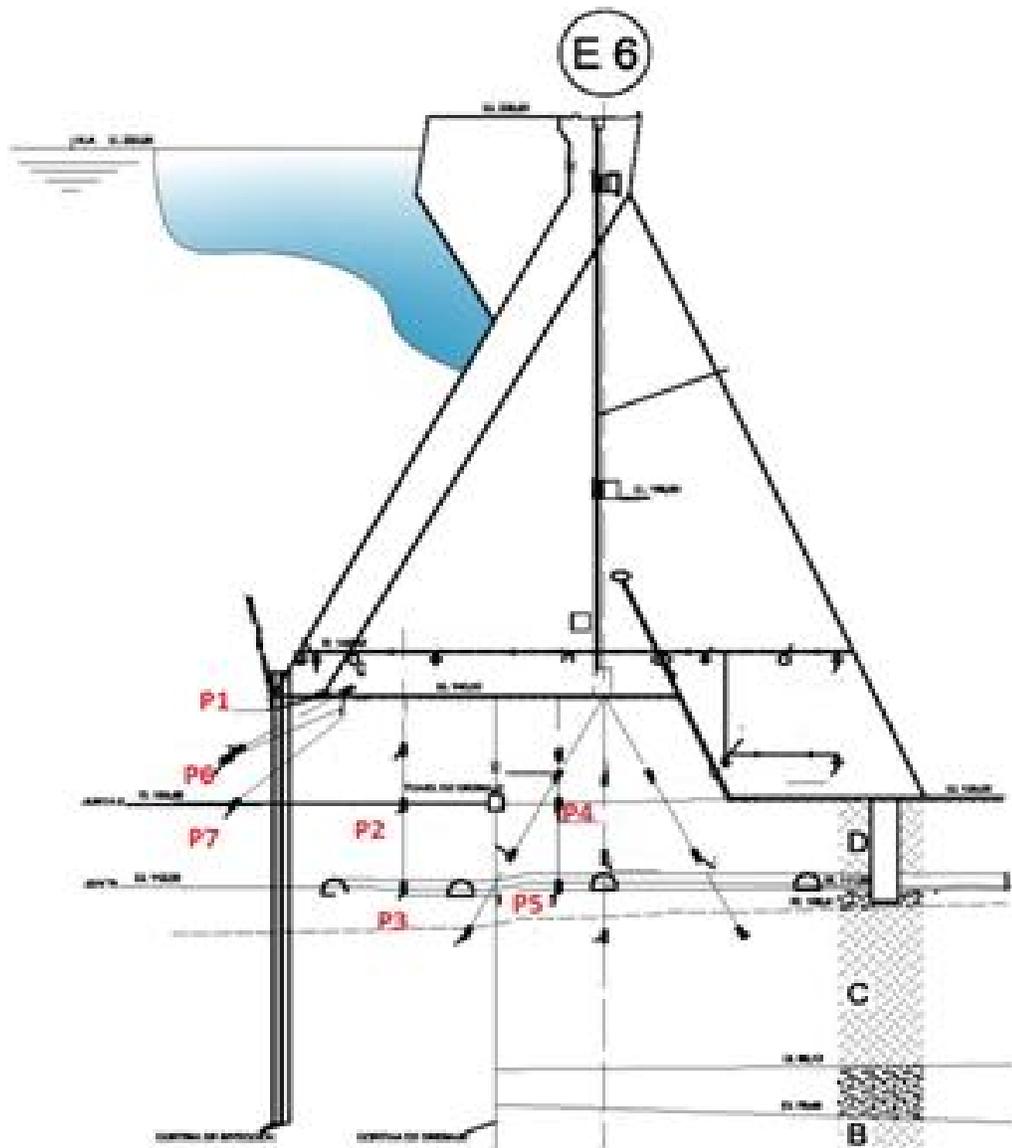


FIGURA 3.8 - Seção do bloco 6 do trecho E de Itaipu com instrumentos p1 até p7 em destaque
Fonte: O autor, adaptado de Itaipu Binacional

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 já foi descrita na seção 2.1.4.

Com o objetivo de 'reduzir' os dados e obter informações sobre relacionamentos não percebidos, foram calculadas as componentes principais do conjunto de dados da fase I. As componentes retidas foram então analisadas quanto a 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 $p < 0,05$).

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 sequência para construção do gráfico de controle multivariado T^2 . Para a fase I, foi construído o gráfico T^2 com a suposição de normalidade multivariada (MVN) dos dados usando a matriz de covariância usual e de diferenças sucessivas (equações 2.1.33 e 2.1.36).

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 3.9.

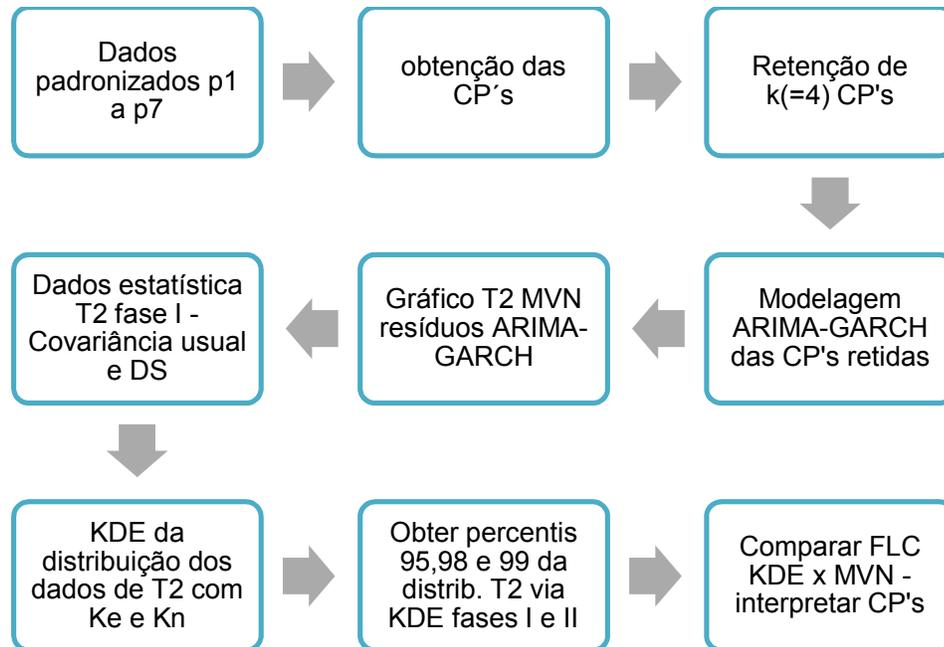


FIGURA 3.9 - Fluxograma do método empregado na seção 3.2.2.2
 Fonte: O autor

Os resultados desta metodologia multivariada estão apresentados na seção 4.2.2 (página 175).

3.2.2.3 Análise de todo conjunto de instrumentos

A última etapa do trabalho propõe uma metodologia multivariada de monitoramento de instrumentos de barragens de concreto, a qual integra as seguintes técnicas: Análise de Componentes Principais (ACP), modelos ARIMAX-GARCH, gráfico de controle multivariado T^2 e a amostragem não paramétrica *Kernel Density Estimation* (KDE). Basicamente, a partir dos dados de um conjunto de instrumentos da barragem, são calculadas, via método ACP, as Componentes Principais (CP's). Em seguida, são produzidas previsões dinâmicas através de modelos ARIMAX-GARCH apenas para as CP's classificadas como "representativas", do ponto de vista da variabilidade total. A análise é realizada a partir dos resíduos de CP's via gráfico de controle multivariado T^2 sem assumir a normalidade, hipótese que pode não ser verdadeira para dados de instrumentação, assim, as densidades preditivas empíricas associadas aos valores da estatística T^2 são geradas, de modo não paramétrico, via KDE.

Finalmente, o limite superior da carta de controle das CP's é obtido via cálculo de percentis e, então, são encontrados os pontos fora do limite de controle do gráfico T^2 . Isto é, é realizado um monitoramento global (realizando as análises dos resíduos das CP's), e não de forma individual para cada instrumento, reduzindo notavelmente o custo operacional de sua execução. A fim de ilustrar a metodologia proposta numericamente, foi feita uma aplicação utilizando dados de instrumentos do Bloco E6 da barragem de concreto da usina hidrelétrica de Itaipu, localizada entre o Brasil e o Paraguai.

Os dados são compostos da média mensal de 383 observações, registradas no período de outubro de 1982 a agosto de 2014. A amostra foi separada, conforme a TABELA 3.1, em amostra de treinamento, para construção e estimativa dos parâmetros dos modelos e amostra de validação, para avaliar o desempenho em termos da capacidade preditiva. Neste caso adotou-se o maior período possível para análise e considerando que houve mudança na periodicidade de leituras, foram adotadas as médias mensais de cada série.

TABELA 3.1 - Separação da amostra

AMOSTRA	FASE	QUANTIDADE DE DADOS	PERÍODO
Treinamento	I	360	10/82 a 09/12
Validação	II	23	10/12 a 08/14
TOTAL		383	

FONTE: O autor

Integram a análise realizada, séries de quatro tipos de instrumentos (totalizando 32 séries), mais a temperatura ambiente com *lags* de retardo de 0 a 11 e o nível do reservatório (ou lago) com *lags* de retardo de 0 a 9, representados no QUADRO 3.3. A temperatura ambiente e o nível do lago são utilizados como variáveis exógenas (explicativas) nos modelos ARIMAX-GARCH das CP's. Estas séries representam a quase totalidade dos instrumentos localizados no trecho E6 da barragem de Itaipu. Portanto, a matriz de dados utilizado consiste de 383 linhas e 54 colunas, correspondentes as 32 séries de instrumentos, mais 12 séries de temperatura ambiente e mais 10 séries do nível do lago.

QUADRO 3.3 – Séries temporais utilizadas no trabalho da seção 3.2.2.3

INSTRUMENTO	CÓDIGO	INSTRUMENTO	CÓDIGO	INSTRUMENTO	CÓDIGO	INSTRUMENTO	CÓDIGO
	1 – A		3A ₁		3D ₁		4A
	1 – B		3A ₂		3D ₂		4B ₁
	1 – C	3 – Base de	3B ₁	3 – Base de	3E ₁		4B ₂
	1 – D	Alongômetro	3B ₂	Alongômetro	3E ₂		4B ₃
1- Piezômetro	1 – E		3C ₁		3F ₁	4 -	4C ₁
	1 – F		3C ₂		3F ₂	Extensômetro	4C ₂
	1 – G						4C ₃
	1 – H						4D ₁
2 – Pêndulo	2 _x	Temperatura	ta, ta1,	Nível do	l ₀ , l ₁ ,		4D ₂
Direto	2 _y	Ambiente	..., ta11	Lago	..., l ₉		4D ₃

FONTE: O autor

As etapas do método proposto são as seguintes.

- Padronização das n variáveis

A razão para tal procedimento já foi descrita na seção 2.1.4.

- Análise e retenção de k componentes principais (ACP) dos dados padronizados das n variáveis

Conforme já mencionado na Seção 2.1.4, a ACP tem o objetivo de gerar poucas variáveis não correlacionadas com a capacidade de resumir a análise dos dados, indicar padrões e variáveis que podem ser monitoradas conjuntamente. Segundo Jackson (1985), há uma série de critérios para a escolha do número de CP's a serem tomadas. Eles incluem: a) testes de significância para a igualdade de raízes, b) retenção de uma quantidade de CP's suficientes para explicar uma proporção determinada da variância total e c) parada quando as variâncias residuais forem iguais a alguma quantidade determinada, usualmente a variabilidade inerente do sistema em estudo.

- Modelagem ARIMAX-GARCH dos valores das CP's extraídas

Este procedimento é útil uma vez que as CP's podem se comportar como séries temporais com estrutura de autodependência, volatilidade e dependentes das

variáveis exógenas aqui consideradas, a saber, nível do reservatório e temperatura ambiente.

Para a seleção de modelos de regressão linear dinâmica das CP's retidas z_i levou-se em conta termos autoregressivos, de médias móveis, temperatura ambiente e nível do lago que fossem estatisticamente significativos e que não houvesse autocorrelação residual significativa ao nível de 95% de confiança.

- Seleção dos melhores modelos ARIMAX-GARCH

Modelos satisfazendo as condições descritas no item anterior foram selecionados para cada CP retida. A seleção do modelo que irá representar cada CP deu-se em função do desempenho da previsão dinâmica de cada um, avaliada no período de estimativa (dentro), de validação (fora) e total da amostra. Foi selecionado em cada CP o modelo que apresentava o melhor desempenho do erro absoluto médio (MAE) e da raiz do erro quadrático médio (RMSE) para a previsão dinâmica, sem inserir os dados reais.

- Obtenção dos resíduos das CP's retidas

Mais uma vez, conforme Montgomery (2013) e Ryan (2011), como os dados de instrumentos podem apresentar (auto)correlação, sazonalidades, volatilidade e dependência de variáveis exógenas que são consideradas causas especiais de variação, o gráfico multivariado T^2 deve ser construído sobre os resíduos dos modelos das CP's retidas.

- Obtenção dos valores da estatística T^2 dos resíduos das CP's e das CP's

Para a construção do gráfico multivariado T^2 que tem o objetivo de avaliar todas as variáveis num único gráfico, foram obtidas as séries de resíduos das k CP's retidas, denominadas res_cp_i , $i = 1, \dots, k$. Estas, bem como, as séries temporais das CP's z_i , $i = 1, \dots, k$ são avaliadas quanto à normalidade (teste Jarque-Bera).

Desta forma, são obtidas duas séries temporais de valores da estatística T^2 , a série oriunda dos resíduos das CP's, denominada t^2_res e a série oriunda das CP's originais, denominada t^2_cps . Para ambas, os dados da fase I, são usados para obter os valores da estatística T^2 com a suposição de normalidade multivariada (MVN) usando as equações (3.2.1), (3.2.2) e (3.2.3)

$$\bar{X} = \sum_{i=1}^m X_i, \quad (3.2.1)$$

$$S = \frac{1}{m-1} \sum_{i=1}^m (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})', \quad (3.2.2)$$

$$T^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' [S]^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}), \quad (3.2.3)$$

onde $\bar{\mathbf{X}}$ é a média amostral, S é a matriz de covariância estimada das m observações e T^2 é a estatística T^2 de Hotelling e foi obtido o limite superior controle (LSC), dado em (2.1.31), para a taxa de erro tipo I, fixado em $\alpha = 1$, $\alpha = 2$ e $\alpha = 5\%$. Do mesmo modo, são obtidos os valores da estatística T^2 e o LSC, dado em (2.1.30), para os dados da fase II.

- Construção de uma estimativa empírica da densidade via KDE

A hipótese de normalidade multivariada pode não ser verdadeira, fazendo com que o LSC das fases I e II gerem muitos falsos alarmes (MASON e YOUNG, 2002). Desta forma, os valores da estatística T^2 na fase I foram usados para a construção de uma densidade de probabilidades empírica via *KDE*.

A escolha desta técnica não paramétrica em particular deu-se em função de que em um estudo comparativo da taxa de falsos alarmes quando o limite superior de controle da estatística T^2 quando determinado com a técnica de reamostragem *Bootstrap*, KDE e distribuição F , equação (2.1.30), realizado por Phaladiganon *et al.* (2013), mostrou que, para amostras geradas da distribuição normal, t e Gama multivariada, a taxa de falsos alarmes via *Bootstrap* e KDE é praticamente igual para todas distribuições e tem desempenho bastante superior ao da distribuição F quando a distribuição não é normal multivariada.

Para os núcleos das equações (2.1.116) e (2.1.117), usam-se $N = 200$ ou 300 na equação (2.1.114) e dois tamanhos para o parâmetro h da estimativa, um padrão fornecido pelo *software Eviews* e outro sendo $h/2$, para avaliar diferentes cenários. O LSC para todos estes cenários ao nível de erro tipo I, fixado em $\alpha = 1$, $\alpha = 2$ e $\alpha = 5\%$ foi determinado para os dados da fase I e da fase II, ou seja, foram obtidos os percentis de 95, 98 e 99%.

A seguir, é avaliada e comparada a quantidade de observações fora do limite de controle (FLC), para ambas as fases, das séries temporais t^2_{res} e t^2_{cps} . Também, busca-se dentre diversas distribuições paramétricas plausíveis a que melhor (em termos de verossimilhança) se ajusta aos dados destas séries.

- Avaliação e comparação da taxa de falsos alarmes

Por fim, é feita uma comparação da taxa de falsos alarmes obtida para o gráfico T^2 considerando o limite superior de controle do gráfico, supondo normalidade multivariada, determinando uma distribuição paramétrica mais adequada aos valores da estatística T^2 e construindo distribuições não paramétricas via KDE.

A FIGURA 3.10 apresenta um fluxograma que resume as etapas da metodologia empregada.

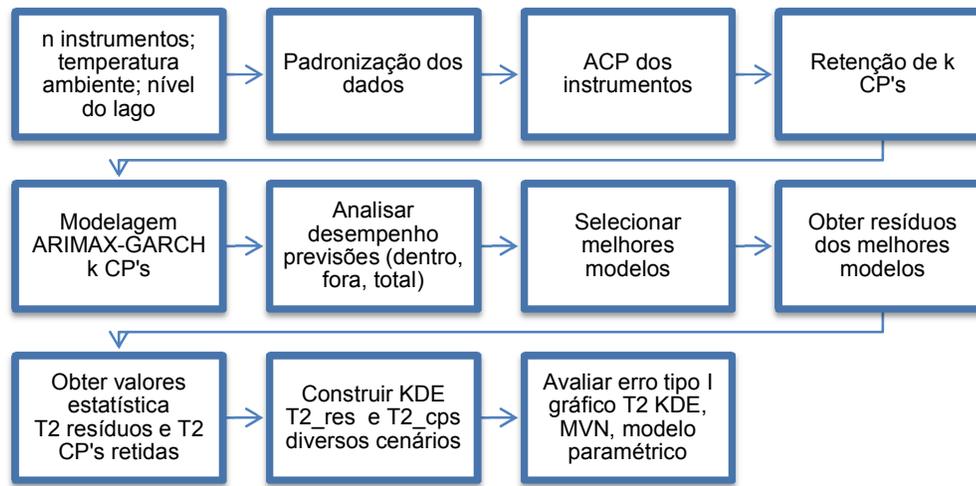


FIGURA 3.10 - Fluxograma das etapas da metodologia para a subseção 3.2.2.3

FONTE: O autor

Os resultados desta metodologia multivariada estão apresentados na seção 4.2.3 (página 185).

4 RESULTADOS

Neste capítulo são detalhados os principais resultados obtidos com a metodologia proposta para tratar com séries temporais de leituras de instrumentos de monitoramento da barragem da Usina Hidrelétrica de Itaipu. Não obstante, os métodos empregados podem ser aplicados em dados de outras barragens. Uma vez que se trabalhou com a abordagem univariada e multivariada, por conveniência, os resultados são separados para cada tipo de tratamento.

Nas seções 4.1 e 4.2, a seguir, são detalhados os principais resultados obtidos que resultaram publicações em revistas ou eventos, cujos resumos e referências podem ser encontradas nos Apêndices de A a H.

4.1 CASO UNIVARIADO

4.1.1 O Caso do Pêndulo Direto

Os resultados desta seção referem-se ao método descrito na subseção 3.2.1.1 (página 145). Inicialmente a série original foi decomposta, segundo a base de *Haar*, nas componentes *Wavelet* (CW) A_1 , D_2 e D_3 , a partir do qual obteve-se o gráfico representado na FIGURA 4.1, onde a variação no eixo y é dada em *mm*.

Em relação a modelagem *ARIMA* da série original (abordagem 1) o melhor modelo encontrado consiste de uma diferenciação estacional, uma sazonal de ordem 12, com os termos autorregressivos de ordem 1, 2 e 4 e sazonal de ordem 12 e de médias móveis de ordem 1 e sazonal de ordem 12.

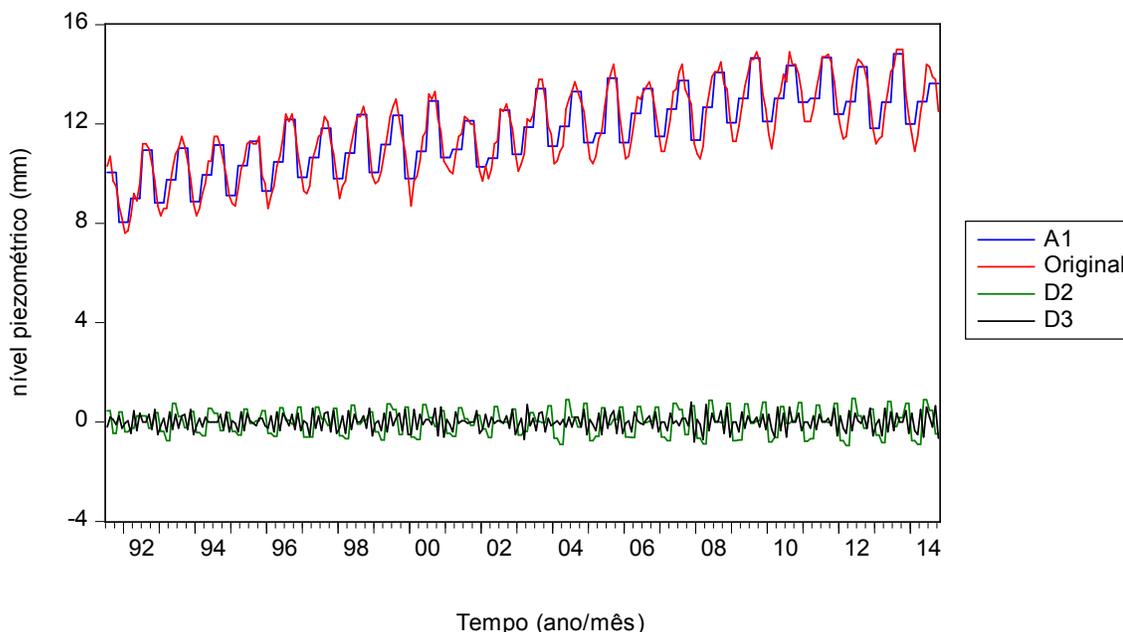


FIGURA 4.1 - Gráficos da série original e suas componentes *Wavelet*
FONTE: Dados da pesquisa – saída do software *EViews*

Foram ajustados modelos de Redes Neurais Artificiais *feedforward* para a série original, resíduos do modelo *ARIMA* da série original, componentes *Wavelet* A_1 , D_2 e D_3 , resíduos da modelagem *ARIMA-GARCH* das componentes *Wavelet*, denominadas séries res_A_1 , res_D_2 e res_D_3 , com janela de previsão (J) variando de 1 a 13 e número de neurônios (N) na camada oculta variando de 1 a 10. Na TABELA 4.1, estão as configurações ótimas encontradas, em qual abordagem foram utilizadas as redes e o valor do *MSE* (definido no final da seção 2.1.5).

TABELA 4.1 - RNA' s ótimas encontradas para as abordagens estabelecidas

SÉRIE	ABORDAGEM	J	N	MSE	SÉRIE	ABORDAGEM	J	N	MSE
<i>Original</i>	2,3	12	6	0,0916	D_3	6,7,8,9	13	7	0,0134
<i>Res_ARIMA</i>	4	6	9	0,0619	res_A_1	10,11	12	10	0,0141
A_1	6,7,8,9	10	6	0,0201	res_D_2	10,11	8	10	0,0029
D_2	6,7,8,9	13	4	0,0039	res_D_3	10,11	12	5	0,0147

FONTE: O autor – saída do software *MatLab*

Para a modelagem *ARIMA-GARCH* das componentes *Wavelet* A_1 , D_2 e D_3 os melhores modelos encontrados estão descritos na TABELA 4.2, onde *DE* indica ordem de diferenciação estacional, *DS* indica que houve diferenciação sazonal de ordem 12, *AR* coeficientes autorregressivos, *MA* coeficientes de médias móveis, *SAR*

coeficientes autorregressivos sazonais, SMA coeficientes de médias móveis sazonais, r e s as respectivas ordens *GARCH*, conforme equação (2.1.86) e a distribuição de referência dos choques aleatórios do modelo.

TABELA 4.2 - Modelos ARIMA-GARCH ótimos encontrados para as componentes

Série	DE	DS	AR	MA	SAR	SMA	r	s	Distribuição
A_1	1	Sim	4,20	4,8	12	–	0	3	t_8
D_2	1	Sim	4,8,16	2,4	12	12	0	2	t_{49}
D_3	0	Sim	–	1	–	12	1	1	t_{19}

FONTE: O autor – saída do software EViews

Na TABELA 4.3 são apresentados os valores das medidas de desempenho, testadas no período de validação da amostra, nas abordagens testadas. Observa-se, em geral, melhora no desempenho da previsão estática com o uso da decomposição *Wavelet* e da combinação de métodos.

TABELA 4.3 - Resultados finais do desempenho das previsões por abordagem

ABORDAGEM	1	2	3	4	5	6	7	8	9	10	11
<i>MAPE</i> (%)	3,13	1,93	2,41	1,65	1,42	1,00	0,97	0,95	0,96	1,17	1,22
<i>MSE</i>	0,2649	0,1176	0,1569	0,0952	0,1098	0,0367	0,0372	0,0351	0,0357	0,0504	0,0608
<i>MAE</i>	0,4005	0,2525	0,3102	0,2163	0,1841	0,1281	0,1239	0,1199	0,1217	0,1530	0,1592
R^2 (%)	0,9805	0,9880	0,9851	0,9898	0,9912	0,9938	0,9940	0,9941	0,9941	0,9927	0,9924

FONTE: O autor, resultados da pesquisa

Na FIGURA 4.2 é exibido o gráfico no período de validação da série original e da melhor abordagem segundo todas as medidas de desempenho, conforme a TABELA 4.3.

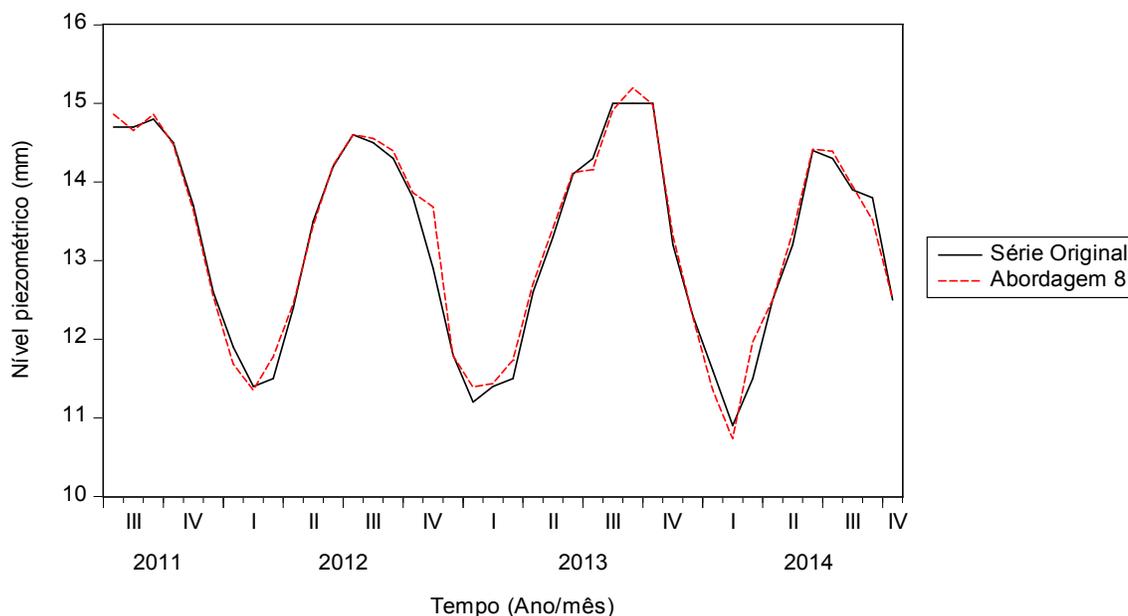


FIGURA 4.2 - Desempenho da melhor abordagem no período de validação

FONTE: O autor, saída do software *EViews*

Os modelos ‘puros’, que constam nas abordagens 1 e 2 do QUADRO 3.2, apresentam o pior desempenho nas previsões, conforme a TABELA 4.3, desta forma o método híbrido que combina estas duas abordagens, na abordagem 3, também não apresenta bons resultados. Já o modelo híbrido sequencial, abordagem 4, consegue uma melhoria nas previsões comparada aos seus antecessores. Todas estas abordagens (1 a 4) têm em comum a não utilização da decomposição *Wavelet*, ou seja, a TABELA 4.3, mostra que todas as abordagens que contemplam a decomposição *Wavelet* (5 a 11) apresentam melhor desempenho de previsão.

Das abordagens com decomposição *Wavelet* (5 a 11), as abordagens 5 e 6 que só levam em conta modelagem *ARIMA-GARCH* ou RNA apresentam desempenho inferior às abordagens híbridas 7, 8 e 9. Isto confirma que, dentre as abordagens híbridas, pós-decomposição *Wavelet*, os modelos híbridos melhoram a previsão. As abordagens 5 e 6 ainda superaram os modelos híbridos 10 e 11, possivelmente, porque estas modelavam com Redes Neurais aos resíduos do modelo *ARIMA-GARCH* e esta estratégia nem sempre produz melhorias.

As abordagens que apresentaram os melhores desempenhos foram 7, 8 e 9. Estas têm em comum, além da decomposição *Wavelet*, o uso de combinações

lineares híbridas dos modelos *ARIMA-GARCH* e de RNA's, combinadas em paralelo, ou seja, quando os modelos *ARIMA-GARCH* e de RNA's são desenvolvidos independentes e combinados linearmente ao final.

Foram construídas onze abordagens de comparação nos quais se variava a utilização ou não de decomposição *Wavelet*, modelagem *ARIMA-GARCH*, com Redes Neurais e combinações híbridas destas técnicas. O melhor desempenho, para todo o período de validação, foi apresentado pela abordagem 8, decomposição *Wavelet* da série original seguida de um modelo híbrido, em paralelo, da soma das combinações ótimas *ARIMA-GARCH* com RNA em cada componente.

Para a série temporal estudada, de modo geral, o uso da decomposição *Wavelet* teve desempenho superior de previsão a não utilização de *Wavelets*, a combinação do uso das técnicas *ARIMA-GARCH* e de RNA's teve desempenho superior ao uso de somente uma delas e as abordagens com melhor desempenho são aquelas que consideram a combinação de todas as técnicas. As combinações híbridas nem sempre tem desempenho superior, isto depende dos desempenhos individuais de cada técnica.

4.1.2 O Caso da Série da Vazão de Afluentes no Rio Paraná

Os resultados desta seção referem-se ao método descrito na subseção 3.2.1.2 (página 149). A FIGURA 4.3 apresenta o gráfico da série temporal completa da vazão na usina de Itaipu, constituída de 492 observações. Foi escolhido o nível de decomposição dois ($p=2$) para preservar a parcimônia no número de modelos individuais a ser estimado. Em relação à base ortonormal, foi escolhida a base de Haar por fornecer as melhores previsões.

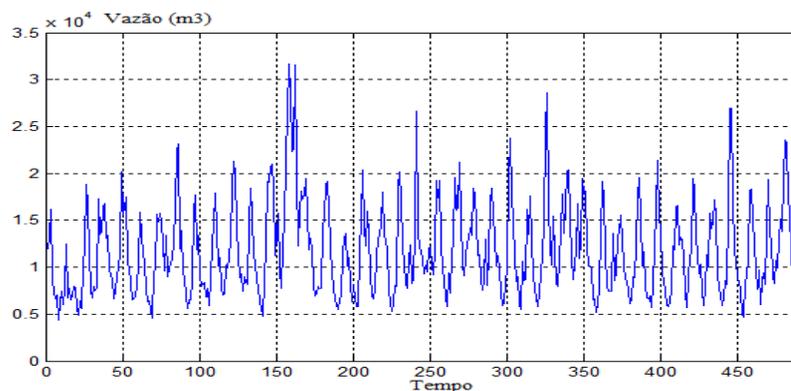


FIGURA 4.3 - Série temporal da vazão média mensal em Itaipu
FONTES: O autor

Na TABELA 4.4 e na TABELA 4.5 nota-se que CL1 foi superior aos métodos preditivos RNA e BJ individualmente, onde os pesos adaptativos lineares ótimos associados com RNA e BJ foram, respectivamente, iguais a 0,618388907 e 0,323249491. O mesmo pode ser verificado usando o método de combinação CL2, quando comparado com os métodos baseados em RNA-WAVELET e BJ-WAVELET. Os pesos (ou coeficientes) lineares ótimos associados a previsões derivadas de RNA-WAVELET e BJ-WAVELET foram, respectivamente, iguais a 0,910473771 e 0,089013323.

TABELA 4.4 - Valores do MAPE nas amostras de treinamento, validação e teste

MÉTODO	MAPE(%)		
	Treinamento	Validação	Teste
RNA	16,35	15,19	21,78
BJ	16,60	19,46	21,25
CL 1	14,95	14,82	19,13
RNA-WAVELET	2,11	2,47	2,80
BJ-WAVELET	2,74	3,13	3,63
CL 2	2,06	2,38	2,64
Método Proposto	1,27	1,27	1,36

Fonte: O autor

TABELA 4.5 - Valores do MAE nas amostras de treinamento, validação e teste

MÉTODO	MAE		
	Treinamento	Validação	Teste
RNA	1.886,87	1.682,38	2.389,97
BJ	2.032,22	2.185,50	2.560,19
CL 1	1.877,10	1.752,25	2.274,33
RNA-WAVELET	224,00	232,73	290,79
BJ-WAVELET	291,92	278,34	396,26
CL 2	218,29	222,45	275,75
Método Proposto	131,83	121,71	141,95

Fonte: O autor

Na FIGURA 4.4 é feita uma comparação entre o método proposto e a abordagem CL2 (que era o melhor na TABELA 4.4 entre os métodos de referência). Note-se que, em todos os 49 meses da amostra de teste, o presente método mostrou-se estável ao longo do tempo para valores do APE (erro absoluto percentual) o que não ocorre para o método CL2 (por exemplo, em 6 meses os valores do APE de CL2 violam o intervalo de 6%, enquanto isto não ocorre em nenhum momento no método proposto). Além disso, na maioria dos instantes, o APE da combinação proposta apresenta valores inferiores aos valores do APE do método CL2.

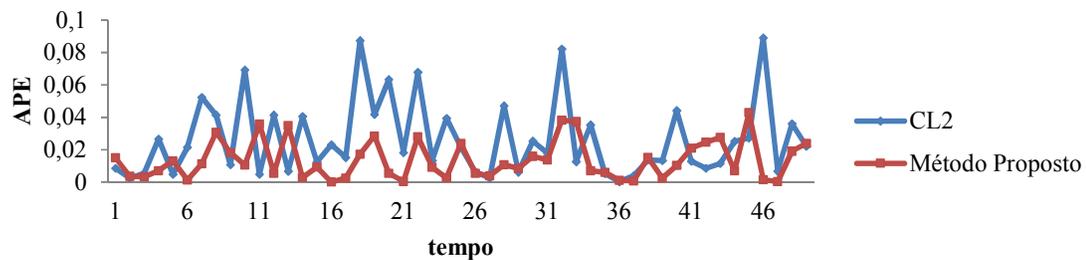


FIGURA 4.4 - Evolução mensal do erro absoluto percentual (APE) na amostra de teste do modelo CL2 (linha azul) e do modelo proposto (linha vermelha)
Fonte: O autor

4.2 CASO MULTIVARIADO

4.2.1 Estatísticas T^2 e Q no Estudo dos Piezômetros

Os resultados apresentados nesta seção referem-se à metodologia detalhada na subseção 3.2.2.1 (página 151). Inicialmente, cada variável (leitura de piezômetro) foi padronizada, ou seja, os dados foram subtraídos de sua média e divididos pela sua variância amostral. Segundo Johnson e Wichern (2007), quando se padronizam as variáveis e trabalha-se com a matriz de correlação ao invés da matriz de covariância, os respectivos pares de autovalores e autovetores, em geral, não são os mesmos, nem mesmo existe uma função simples que os relaciona. Sugere-se que variáveis que estão em escalas de domínio e variação muito diferentes, devem ser padronizadas. Para Montgomery (2013), 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.

Para avaliar a hipótese de normalidade multivariada foi usado um teste prático de ajustamento de χ_p^2 descrito em Mingoti (2005) e Johnson e Wichern (2007). Para cada vetor $\mathbf{X}' = (X_1, X_2, \dots, X_7)$ contendo as leituras padronizadas dos piezômetros foi calculado se

$$(\mathbf{X} - \bar{\mathbf{X}})'S^{-1}(\mathbf{X} - \bar{\mathbf{X}}) \leq \chi_{7;95\%}^2, \quad (4.2.1)$$

onde $\chi_{7;95\%}^2$ é o valor que corresponde ao percentil de 95% da distribuição χ_7^2 . Uma vez que exatamente 95% da amostra da primeira fase satisfaz a condição, pode-se aceitar a hipótese que os dados provêm de uma distribuição normal multivariada ao nível de 95% de confiança.

Para efeito de comparação, gráficos univariados de controle de Shewhart da média amostral \bar{x} foram construídos para a fase I, considerando o $LSC = \bar{x} + 3\hat{\sigma}$, $LIC = \bar{x} - 3\hat{\sigma}$ e $\hat{\sigma}$ é, conforme Montgomery (2013), um estimador não viesado para σ , dado por $\hat{\sigma} = \bar{R}/d_2$, onde \bar{R} é a amplitude média e $d_2 = 1,128$ é um valor tabelado em função do tamanho da amostra, de modo que a taxa de falsos alarmes seja $\alpha = 0,27\%$. Na TABELA 4.6 é apresentada a quantidade de observações fora dos limites de controle (FLC) para os respectivos gráficos univariados de Shewhart dos instrumentos P_1 a P_7 aos 300 dados da fase I. Observa-se a inviabilidade do tratamento univariado com a elevada quantidade de observações fora dos limites de controle.

TABELA 4.6 - Observações fora dos limites de controle (FLC) para gráficos univariados

Inst.	FLC												
P_1	239	P_2	178	P_3	203	P_4	82	P_5	149	P_6	190	P_7	166

Fonte: O autor

Na seleção do estimador da matriz de covariância a ser usada no gráfico T^2 para os dados dos sete piezômetros padronizados foram testadas a matriz de covariância usual (2.1.33) e a matriz de diferenças sucessivas (2.1.36). Na TABELA 4.7 é apresentado o número de observações fora do limite de controle (FLC) para o gráfico T^2 para cada tipo de matriz de covariância testada, para uma probabilidade de erro tipo I fixada em $\alpha = 0,0027$ para cada variável, que corresponde aos limites 3σ dos gráficos de Shewhart, então, segundo Montgomery (2013), a verdadeira probabilidade de erro tipo I, caso as variáveis sejam independentes, para o

procedimento de controle conjunto é $\alpha' = 1 - (1 - \alpha)^7 = 0,0187$. Deve-se destacar, é claro, que as variáveis originais não são independentes.

Não obstante, o tratamento multivariado passa a ser viável, uma vez que a probabilidade de que 8 ou mais observações, num universo de 300, estejam, ao acaso, acima do $LSC_{T^2}^I$ se os dados provém de uma distribuição normal multivariada é de 0,2026, para $\alpha' = 0,0187$, não se pode rejeitar a hipótese de que o processo está sob controle estatístico ao nível de 95% de confiança. Como a matriz de diferenças sucessivas é mais sensível a pequenos desvios da média, para a proposta deste trabalho, foi selecionada a matriz de covariância usual.

TABELA 4.7 - Número de observações FLC para cada matriz de covariância para o gráfico T^2 .

MATRIZ	FLC	OBSERVAÇÕES	LSC
Usual	8	25,68,249,250,264,268,269,270	16,53
Diferenças sucessivas	300	Todas	

Fonte: O autor

Extraindo as componentes principais do conjunto de dados da fase I, foi realizada uma simulação caso sejam selecionadas $k = 2,3,4,5,6$ componentes principais, que explicam um percentual da variabilidade, conforme a TABELA 4.8. Os autovetores que constituem a matriz U da decomposição espectral de S são exibidos na TABELA 4.9. Os dados dos escores das k componentes principais foram retidos e foi testada a normalidade multivariada pelo teste χ_k^2 ao nível de confiança de 95% do que se obteve a aceitação da hipótese de normalidade, conforme a segunda coluna da TABELA 4.10.

Por exemplo, o limite superior de controle para a estatística T_2^2 (escores das 2 componentes principais) na fase I foi calculado pela equação (2.1.31) e resultou em $LSC_{T^2}^I = 10,30$ e todas as 300 observações desta fase não ultrapassaram o LSC . Para a fase II, o limite superior de controle da estatística T_2^2 calculado pela equação (2.1.30) resultou em $LSC_{T^2}^{II} = 10,44$. Para os demais números de CP 's o LSC está na quarta e oitava colunas da TABELA 4.10. Na quinta e nona colunas da TABELA 4.10 estão a quantidade de observações $FLC_{T^2}^I$ e $FLC_{T^2}^{II}$ e quais são estas observações (OBS) estão na sexta e décima coluna para os gráficos T_k^2 nas fases I(II). Por fim, na sétima e décima primeira colunas da TABELA 4.10, a probabilidade de que a quantidade de

FLC obtida esteja sob controle estatístico ao nível de 95%, ou seja, valores menores do que 5% devem rejeitar a hipótese de que os dados estão sob controle.

O gráfico de controle da estatística T_4^2 compreendendo os dados das duas fases está apresentado na FIGURA 4.5, e na FIGURA 4.6 é exibida a elipse de controle de 95% de confiança dos dados para T_4^2 das duas primeiras componentes principais, pontos em vermelho representam observações fora de controle, que ficam fora da elipse de controle se forem tomadas as componentes 2 e 3. Deve ser salientado que estes gráficos foram construídos para uma taxa de erro tipo I, $\alpha = 0,0027$ e pela matriz de covariância usual, equação (2.1.33), ao nível de confiança de 95%.

Os autovetores da matriz de covariância S formam as colunas da matriz ortogonal U da decomposição espectral de S , de modo que

$$U'SU = L, \quad (4.2.2)$$

onde L é a matriz diagonal dos autovalores que representam a variância de cada componente principal (JACKSON, 1991). Desta forma pode-se transformar p variáveis correlacionadas $X = (x_1, x_2, \dots, x_p)$ em p novas variáveis não correlacionadas $Z = (z_1, z_2, \dots, z_p)$ por meio da transformação

$$Z = U'X. \quad (4.2.3)$$

Também é válido que

$$X = UX. \quad (4.2.4)$$

Entretanto, quando se usa um número $k < p$ de componentes principais, então se toma a sub-matriz de ordem $p \times k$ de U e o sub-vetor de ordem $k \times 1$ no produto UZ , obtendo-se, portanto, uma aproximação para X que será representada por \hat{X} . Se forem usadas $k = 2, \dots, 6$ das $p = 7$ componentes principais, então se toma a sub-matriz de ordem $7 \times k$ de U e o sub-vetor de ordem $k \times 1$ no produto UZ , obtendo-se uma aproximação para X , representada por \hat{X} , por meio das equações (4.2.2), (4.2.3) e (4.2.4). O valor obtido para o *LSC* da estatística Q , usando as equações (2.2.7), (2.2.8) e (2.2.6) foi $Q_\alpha = 9,95$ quando $k = 2$, os demais valores de Q_α para $k = 3, \dots, 6$ estão na segunda coluna da TABELA 4.11. As demais colunas da TABELA 4.11 apresentam a quantidade de FLC_I e FLC_{II} e quais são estas observações(*OBS*).

A FIGURA 4.7 apresenta o comportamento da estatística Q ao longo do período de análise, para $k = 4$, calculada pela equação (2.1.44).

TABELA 4.8 - Variabilidade Explicada pelas componentes principais

COMPONENTE	AUTOVALOR	VARIABILIDADE	VARIABILIDADE
		EXPLICADA (%)	ACUMULADA (%)
1	3,55585	50,80	50,80
2	1,45125	20,73	71,53
3	1,02258	14,61	86,14
4	0,50179	7,17	93,31
5	0,26410	3,77	97,08
6	0,13599	1,94	99,02
7	0,06845	0,98	100,00

FONTE: O autor, resultados da pesquisa

TABELA 4.9 - Autovetores da matriz de covariância amostral S (equação (2.1.33))

z_1	z_2	z_3	z_4	z_5	z_6	z_7
-0,1029	-0,6934	-0,3506	0,4859	-0,0651	0,3719	-0,0839
0,3485	-0,4289	0,1239	-0,7007	-0,3266	0,2820	0,0454
0,4481	0,3431	0,1700	0,1886	0,1189	0,5824	-0,5135
0,1458	-0,4160	0,7757	0,2213	0,3095	-0,2426	-0,0161
0,4588	0,1106	0,0659	0,4222	-0,6996	-0,2263	0,2324
0,4456	-0,1774	-0,4145	-0,0887	0,1897	-0,5599	-0,4906
0,4867	0,0274	-0,2342	0,0477	0,5036	-0,1391	0,6574

FONTE: O autor, resultados da pesquisa

TABELA 4.10 - Resultados do Gráfico T^2 das k CP's obtido a partir da retenção de k componentes principais

k	%	α' (%)	FASE I				FASE II			
			$LSC_{T^2}^I$	$FLC_{T^2}^I$	OBS.	Prob.	$LSC_{T^2}^{II}$	$FLC_{T^2}^{II}$	OBS.	Prob.
2	96,67	0,539	10,30	0	-	100%	10,44	0	-	100%
3	96,00	0,808	11,64	5	249,250,263,268,269	9,81%	11,81	7	249,250,268,269,317,318,319	1,6%
4	96,00	1,07	12,91	5	249,250,263,268,269	22,58%	13,12	7	249,250,263,268,269,318,319	6,01%
5	95,67	1,34	14,15	7	68,249,250,263,268,269,270	11,13%	14,37	8	68,249,250,268,269,270,318,319	6,78%
6	94,67	1,61	15,35	7	68,249,250,264,268,269,270	21,21%	15,60	8	68,249,250,264,268,269,270,318	14,63%

FONTE: O autor, resultados da pesquisa

TABELA 4.11 - Resultados do Gráfico Q das p-k componentes principais

k	VALORES E OBSERVAÇÕES				
	Q_α	FLC_I	OBS.	FLC_{II}	OBS.
2	9,95	4	249,250,263,269	4	249,250,263,269
3	4,41	2	25,264	2	25,264
4	2,09	3	25,140,249	3	25,140,249
5	0,96	1	25,138	2	25,138,310
6	0,39	4	25,51,135,170	4	25,51,135,170

FONTE: O autor, resultados da pesquisa

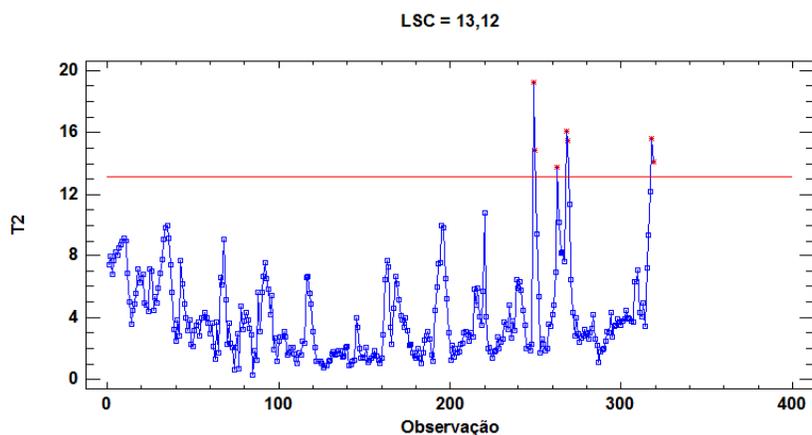


FIGURA 4.5 – Gráfico T^2 para quatro componentes principais retidas incluindo dados das fases I e II
 FONTE: O autor, resultados da pesquisa

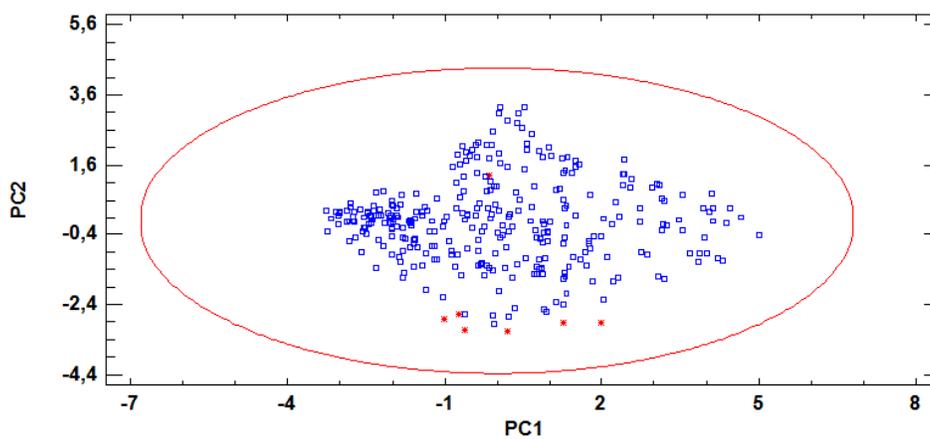


FIGURA 4.6 - Elipse de controle dos escores das duas primeiras componentes de quatro componentes principais retidas
 FONTE: O autor, resultados da pesquisa

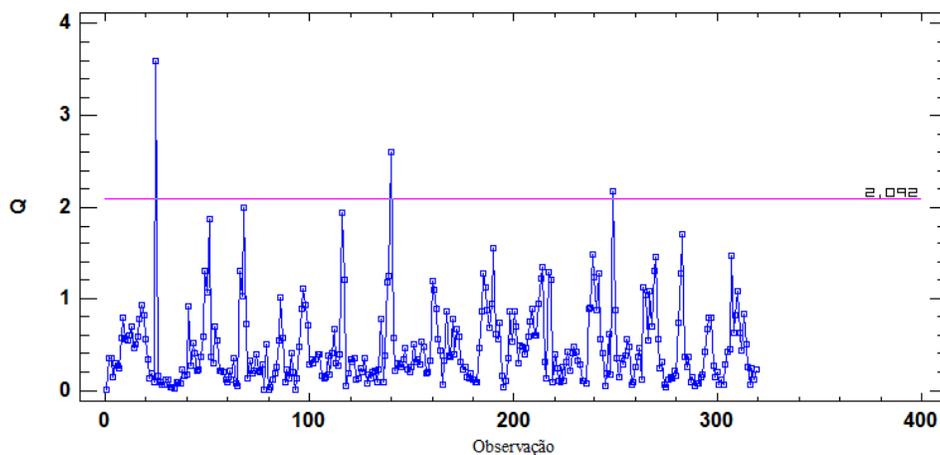


FIGURA 4.7- Gráfico da estatística Q considerando $k=4$ componentes principais nas fases I e II
 FONTE: O autor, resultados da pesquisa

Este trabalho multivariado buscou estabelecer um método para tratar com gráficos de controle para instrumentos de monitoramento de barragens, sete piezômetros, no caso. Na prática, diante da grande quantidade de instrumentos em uma grande barragem e a correlação existente entre eles, o monitoramento individualizado de cada instrumento pode ser inviável, seja pelo excesso de gráficos a analisar ou pela grande quantidade de falsos alarmes emitidos que pode desacreditar o sistema. O método proposto envolve o tratamento multivariado e resume a análise de um conjunto de instrumentos às estatísticas T^2 e Q combinado com *ACP* para explicar, respectivamente, a variabilidade inerente (de causas atribuíveis) e de fontes aleatórias no sistema. O objetivo foi reduzir o trabalho usando análise multivariada, reduzir os falsos alarmes a níveis estatisticamente sob controle e identificar diferenças nas observações fora dos limites de controle para as estatísticas T^2 e Q .

É importante salientar que caso se tivesse optado pela análise de sete gráficos univariados de Shewhart para a média e os limites de controle de $\pm 3\hat{\sigma}$, de cada piezômetro, seriam obtidos entre 82 e 239 valores fora de controle, ou seja, o problema de estabelecer valores limites para a leitura de instrumentos não poderia ser tratado por esta técnica.

Os resultados mostraram que o modelo de componentes principais combinado com a estatística Q conseguiu ajustar melhor os dados da fase I quando são tomadas, ao menos, quatro componentes principais, pois, neste caso, as observações indicadas como fora de controle para o gráfico T^2 aparecem como fora de controle ou para o gráfico T^2 das k componentes principais ou para o gráfico de Q e conseguiu mostrar que, no estudo de caso aplicado, o sistema de monitoramento multivariado dos piezômetros, localizados no trecho E da usina hidrelétrica de Itaipu, está sob controle estatístico ao nível de 95% de confiança, independente do modelo tomado, ou seja, gráfico T^2 ou componentes principais mais a estatística Q .

Outro benefício da utilização combinada destas estatísticas com quatro componentes principais foi que das observações apontadas como fora de controle, pode-se fazer uma distinção entre elas. Entre as observações dadas como fora de controle da estatística Q para quatro *CP's* a observação 25 está relacionada com o máximo global do instrumento p_7 , a observação 140 está associada com um máximo

local de p_5 (ambas são aparentes *outliers*) e a observação 249 está associada com o máximo global de p_4 que é o instrumento com menor variância.

As quatro primeiras componentes principais explicaram mais de 90% da variabilidade. Foi observado que a primeira componente principal possui uma interpretação em termos de uma contribuição média de cada instrumento para a variabilidade global que depende da camada e da elevação em que se encontram.

De acordo com a primeira coluna da TABELA 4.9, pode-se observar que o instrumento p_7 é o que tem maior contribuição para esta componente principal e sua localização é antes de uma cortina de injeção de concreto e numa elevação inferior, ou seja, local teoricamente mais susceptível a subpressões. O segundo instrumento com maior contribuição nesta componente principal é p_5 localizado numa junta com elevação inferior, conforme a FIGURA 3.8. Já o instrumento que quase não exerce influencia nesta componente é p_1 e sua localização é após a cortina de injeção na elevação mais superior dos instrumentos em estudo. Isto confirma a eficiência da cortina de injeção de concreto da obra da barragem.

O modelo adotado permite uma interpretação das componentes principais (estatística T^2) como uma variabilidade consequente de fatores ambientais e inerentes ao modelo. É fundamental compreender que a variabilidade das componentes principais é oriunda destes fatores e também controlar separadamente a variabilidade aleatória que pode estar ligada a *outliers* do modelo (estatística Q).

O uso de componentes principais tem outra vantagem que é superar o problema de singularidades. Neste estudo de caso, por exemplo, o determinante da matriz de covariância de todas as variáveis originais era 0,0065. As singularidades estão associadas com a existência de autovalores quase nulos que podem gerar problemas computacionais na inversão da matriz de covariância e no consequente cálculo da estatística T^2 .

4.2.2 O Caso dos Piezômetros do Trecho E6 via KDE

Os resultados apresentados nesta seção referem-se a metodologia detalhada na subseção 3.2.2.2 (página 152). Considerando o conjunto de dados da fase I, na

TABELA 4.12, TABELA 4.13 e TABELA 4.14 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).

TABELA 4.12 - Análise das componentes principais de p_1 a p_7 autovalores (esquerda) e autovetores (direita)

Comp.	Autovalor	Proporção	Prop. Acum.	Variável	p_1	p_2	p_3	p_4	p_5	p_6	p_7
1	3,5558	0,5080	0,5080	p_1	-0,1029	0,6934	-0,3506	0,4858	-0,0651	0,3719	0,0838
2	1,4512	0,2073	0,7153	p_2	0,3485	0,4288	0,1238	-0,7007	-0,3265	0,2820	-0,0454
3	1,0225	0,1461	0,8614	p_3	0,4480	-0,3430	0,1700	0,1886	0,1189	0,5824	0,5135
4	0,5017	0,0717	0,9331	p_4	0,1457	0,4160	0,7757	0,2213	0,3095	-0,2425	0,0160
5	0,2640	0,0377	0,9708	p_5	0,4587	-0,1106	0,0659	0,4221	-0,6996	-0,2262	-0,2324
6	0,1359	0,0194	0,9902	p_6	0,4455	0,1774	-0,4145	-0,0886	0,1897	-0,5599	0,4906
7	0,0684	0,0098	1,0000	p_7	0,4867	-0,0274	-0,2342	0,0477	0,5036	0,1390	-0,6574

FONTE: O autor, resultados da pesquisa.

TABELA 4.13 - Análise das componentes principais de p_1 a p_7 - Correlações

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

FONTE: O autor, resultados da pesquisa

TABELA 4.14 - Análise das componentes principais de p_1 a p_7 - Teste de normalidade

COMPONENTE	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

FONTE: O autor, resultados da pesquisa

TABELA 4.15 - Teste de normalidade Jarque-Bera aplicado aos dados originais padronizados (p_1 a p_7)

INSTRUMENTO	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

FONTE: O autor, resultados da pesquisa

Observa-se que as CP's 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 4.15) rejeitam a hipótese sob

os instrumentos (valor $p < 0,05$) e para as CP's 2,4,5 e 6 já se pode aceitar a hipótese de normalidade (TABELA 4.14). Foram selecionadas 4 CP's tendo em vista que elas explicam mais de 90% da variabilidade, conforme TABELA 4.12.

A análise dos correlogramas das CP's mostrou que existe forte autocorrelação em todas elas e na FIGURA 4.8 é exibido o correlograma de pc1, o que indica que a série pode ser mais bem modelada via método *ARIMA-GARCH*.

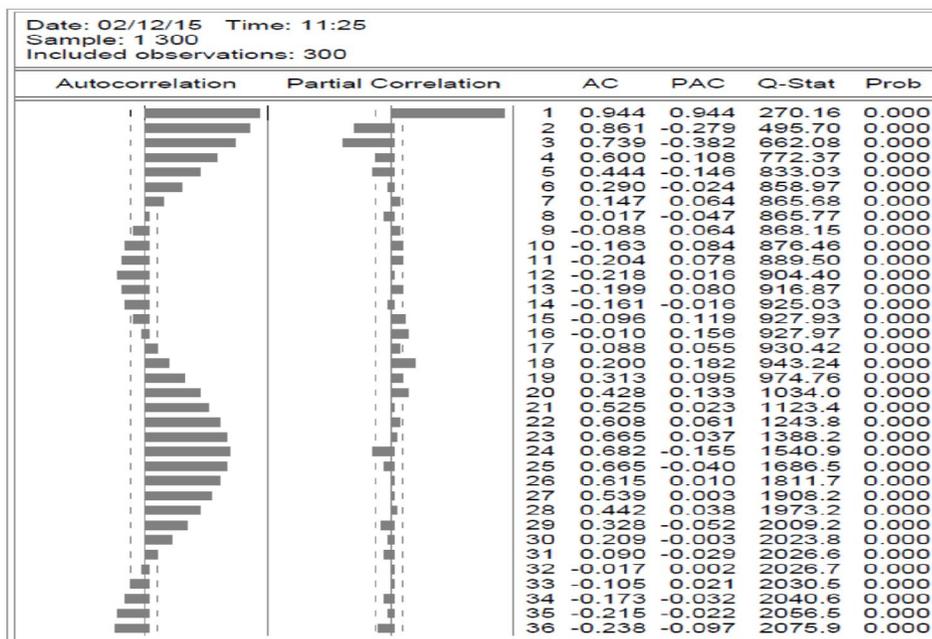


FIGURA 4.8 - Correlograma de pc1 (FAC e FACP)
FONTE: O autor, resultados da pesquisa

A TABELA 4.16 apresenta o resultado da avaliação da significância dos coeficientes *ARIMA-GARCH(1,1)* da variável pc1 e a FIGURA 4.9 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 principais 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*.

TABELA 4.16 - Resultados da modelagem ARIMA-Garch de pc1

Variável	Coefficiente	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 ²	0,93491	Estatística Durbin Watson	1,9387	

FONTE: O autor, resultados da pesquisa

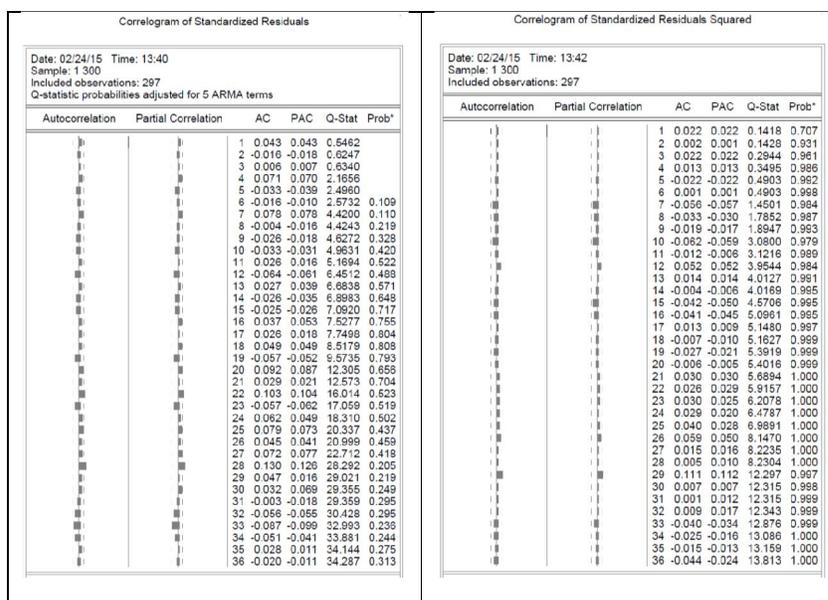


FIGURA 4.9 - Correlograma dos resíduos e resíduos quadráticos de pc1

FONTE: O autor, resultados da pesquisa

A TABELA 4.17 apresenta o resultado do teste de normalidade de Jarque-Bera para normalidade univariada para cada uma das CP's (dados de ambas as fases) e a TABELA 4.18 apresenta o resultado do teste de normalidade de Jarque-Bera para normalidade univariada para as séries dos resíduos de cada CP retida. Nota-se que somente um dos resíduos das CP's apresenta distribuição normal ao nível de significância de 95% e como as distribuições marginais serem normais é condição necessária para a distribuição conjunta das quatro CP's serem normais multivariadas, podemos rejeitar esta hipótese. A tentativa de transformação dos dados para obter normalidade também não surtiu efeito.

TABELA 4.17 - Teste de normalidade Jarque-Bera para as componentes retidas (fases I e II)

Teste/Série	PC1	PC2	PC3	PC4
Jarque-Bera	13.17534	0.861922	7.636870	4.903472
Probabilidade	0.001377	0.649884	0.021962	0.086144

FONTE: O autor, resultados da pesquisa

TABELA 4.18 - Teste de normalidade Jarque-Bera para os resíduos do modelo ARIMA-GARCH das componentes retidas (fases I e II)

Teste/Série	RES_PC1	RES_PC2	RES_PC3	RES_PC4
Jarque-Bera	144.9864	3350.670	550.8130	3.491605
Probabilidade	0.000000	0.000000	0.000000	0.174505

FONTE: O autor, resultados da pesquisa

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 (2.1.30) e (2.1.31) se baseiam numa distribuição F ou β .

Com os dados de referências da estatística T^2 seja ela calculada via matriz de covariância usual ou de diferenças sucessivas equações (2.1.33) e (2.1.36) na fase I, foram construídas distribuições de probabilidade empíricas com *Kernel Density Estimation* usando os núcleos de Epanichinikov e da distribuição normal (equações (2.1.116) e (2.1.117)) 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 CP's retidas (1 a 4). A FIGURA 4.10 contém o histograma de dados de T^2 e duas *Kernel* funções de estimativa da densidade.

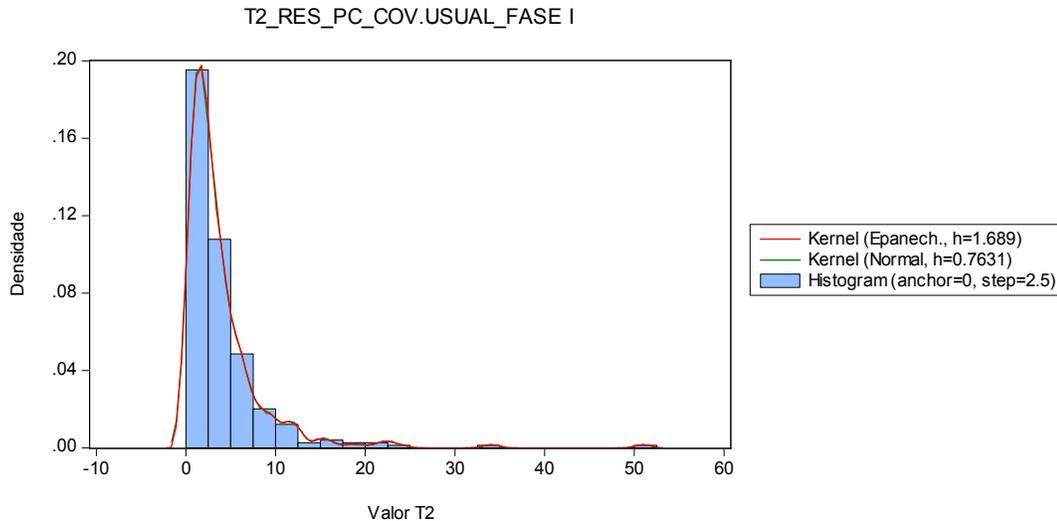


FIGURA 4.10 - Exemplo de kernel funções de estimativa da densidade de probabilidade dos valores de T^2
FONTE: O autor, resultados da pesquisa

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 4.19 (covariância usual) e TABELA 4.20 (covariância por diferenças sucessivas) juntamente com o respectivo número de FLC considerando normalidade multivariada (MVN) e o número médio esperado (CMS).

Observa-se que os resíduos das CP's 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 CP's originais (linhas 13,14 da TABELA 4.19 e TABELA 4.20). Entretanto, ao usar a matriz de covariância estimada pela equação (2.1.36), observa-se que o número de FLC para o gráfico T^2 das CP's 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.

TABELA 4.19 - Comparação de FLC gráfico T^2 matriz de covariância usual KDE (linhas 1-12) x MVN (linhas 13,14)

Covariância Usual	1%			2%			5%		
	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II
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

T2 Resíduos PC1 a PC4 - 297 obs.

FONTE: O autor, resultados da pesquisa

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 três taxas de falsos alarmes fixadas.

TABELA 4.20 - Comparação de FLC gráfico T^2 matriz de covariância por diferenças sucessivas KDE (linhas 1- 12) x MVN (linhas 13,14)

Cov. Dif. Sucessiv.	1%			2%			5%		
	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II	LSC I	FLC Fase I	FLC Fase II
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

T2 Resíduos PC1 a PC4 297 obs.

FONTE: O autor, resultados da pesquisa

Fixado o LSC (22,08 conforme linha 9/coluna 2 da TABELA 4.19) da estatística 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 rigorosa) 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 CP's). O gráfico T^2 deste cenário para a fase I é exibido na FIGURA 4.11 e os valores da contribuição de cada variável (resíduo pc1 a pc4) segundo a equação (2.1.37) estão na TABELA 4.21.

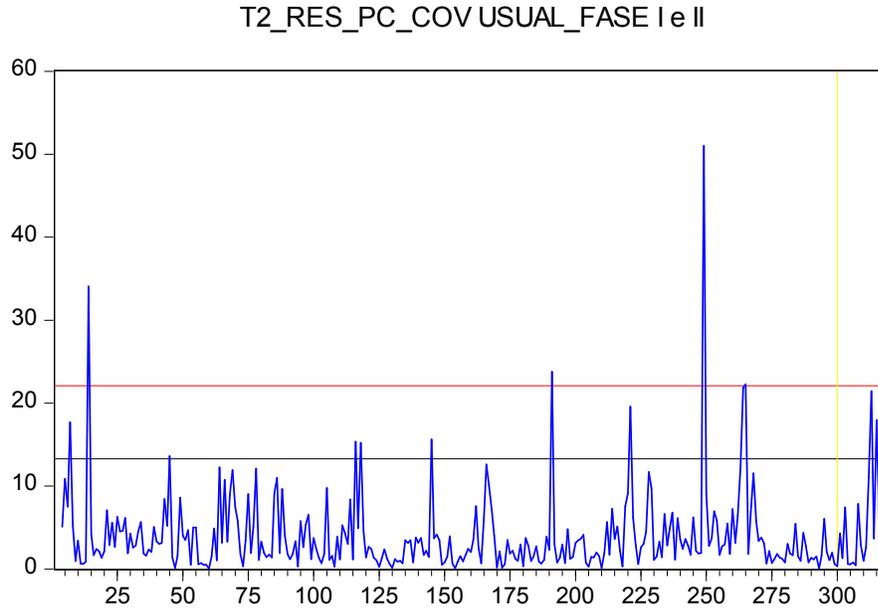


FIGURA 4.11 – Gráfico do escore T^2 dos resíduos das PC's - matriz covariância usual - fases I e II - LSC MVN(preto) e LSC KDE(vermelho) em relação a observação (amostra)
FONTE: O autor, resultados da pesquisa

Da análise do peso da cada variável (piezômetro) nas componentes principais (TABELA 4.12) nota-se que os piezômetros que tem maior influência 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 da maior permeabilidade do derrame basáltico em que se encontram, conforme FIGURA 3.8 (página 154), mais propensa a subpressões.

TABELA 4.21 - Decomposição T^2 - Contribuição relativa para o valor de T^2 - valores de d_i

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

FONTE: O autor, resultados da pesquisa

Por exemplo, na TABELA 4.21, 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 sete variáveis à análise de uma variável (estatística T^2), entretanto, isto não impediu a identificação de quando e onde elas ocorreram.

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.

Ao se buscar 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 (novamente os mesmos piezômetros), 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 normalidade multivariada) 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 normal multivariado.

A combinação de *ACP* e *KDE* se mostrou promissora para estimar percentis 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.

4.2.3 Análise de Todo Conjunto de Instrumentos

Os resultados apresentados nesta seção referem-se à metodologia detalhada na subseção 3.2.2.3 (página 156). A análise de componentes principais (ACP) levou a seleção (retenção) de cinco CP's (z_1, z_2, z_3, z_4, z_5) para posterior construção de modelos ARIMAX-GARCH e do gráfico T^2 . As razões para a escolha de 5 CP's foram o alto percentual de variabilidade explicado (elas explicam quase 90% da variabilidade) e que estas 5 componentes principais (CP's) conseguem incluir todas as variáveis significativamente em ao menos uma (e somente uma) das CP's, ou seja, não se 'perde' o monitoramento de nenhuma variável e torna o modelo mais simplificado possível.

A CP z_1 é principalmente determinada pelos instrumentos: z_x (pêndulo direto medido na direção do fluxo do rio), todos do tipo 3 (A a F – bases de alongômetro), extensômetros 4A e do grupo 4B e piezômetro 1A. Foram selecionados modelos ARIMAX-GARCH que satisfizeram as condições citadas na seção 3.2.2.3 (página 156), para cada uma das CP's retidas. A TABELA 4.22 contém o grau de explicação R^2 , quais os termos ta (temperatura ambiente), $'s'ar$ (sazonal autorregressivo), $'s'ma$ (sazonal de médias móveis), a ordem $GARCH(r,s)$ inseridos nos modelos de z_1 , o valor do teste DW (equação (2.1.77)) e o valor do critério de informação Akaike (AIC). Nenhum coeficiente para o nível do lago foi significativo para z_1 . Foram feitas 5 modelagens para z_1 (z_{1_1} a z_{1_5}). De forma análoga, isto foi realizada para as demais CP's.

TABELA 4.22 - Resumo da modelagem ARIMA-GARCH da CP z_1

MODELO	$z_{1,1}$	$z_{1,2}$	$z_{1,3}$	$z_{1,4}$	$z_{1,5}$
R^2	79,79	99,38	99,38	99,38	99,52
ta	1,7	0 a 10	0 a 10	0 a 10	0 a 10
DW	2,03	2,03	2,04	2,04	2,04
AIC	0,4751	0,3831	0,3890	0,3709	0,3841
$ar(sar)$	$ar(1), sar(12)$	$ar(1), sar(12)$	$ar(1), sar(12)$	$ar(1)sar(12)$	$ar(1)$
$ma(sma)$	--	$ma(1)$	$ma(1)$	$ma(1)sma(24)$	$ma(1)sma(12)$
GARCH	--	--	(0,1)	(0,2)	--

Fonte: O autor

Variáveis com coeficiente negativo em z_1 têm tendência de decrescimento e as demais de crescimento, conforme a FIGURA 4.12, onde o eixo x é o tempo e o eixo y contém a leitura padronizada do instrumento. Estes devem ser considerados os instrumentos mais importantes da análise, porque apresentam maior variabilidade.

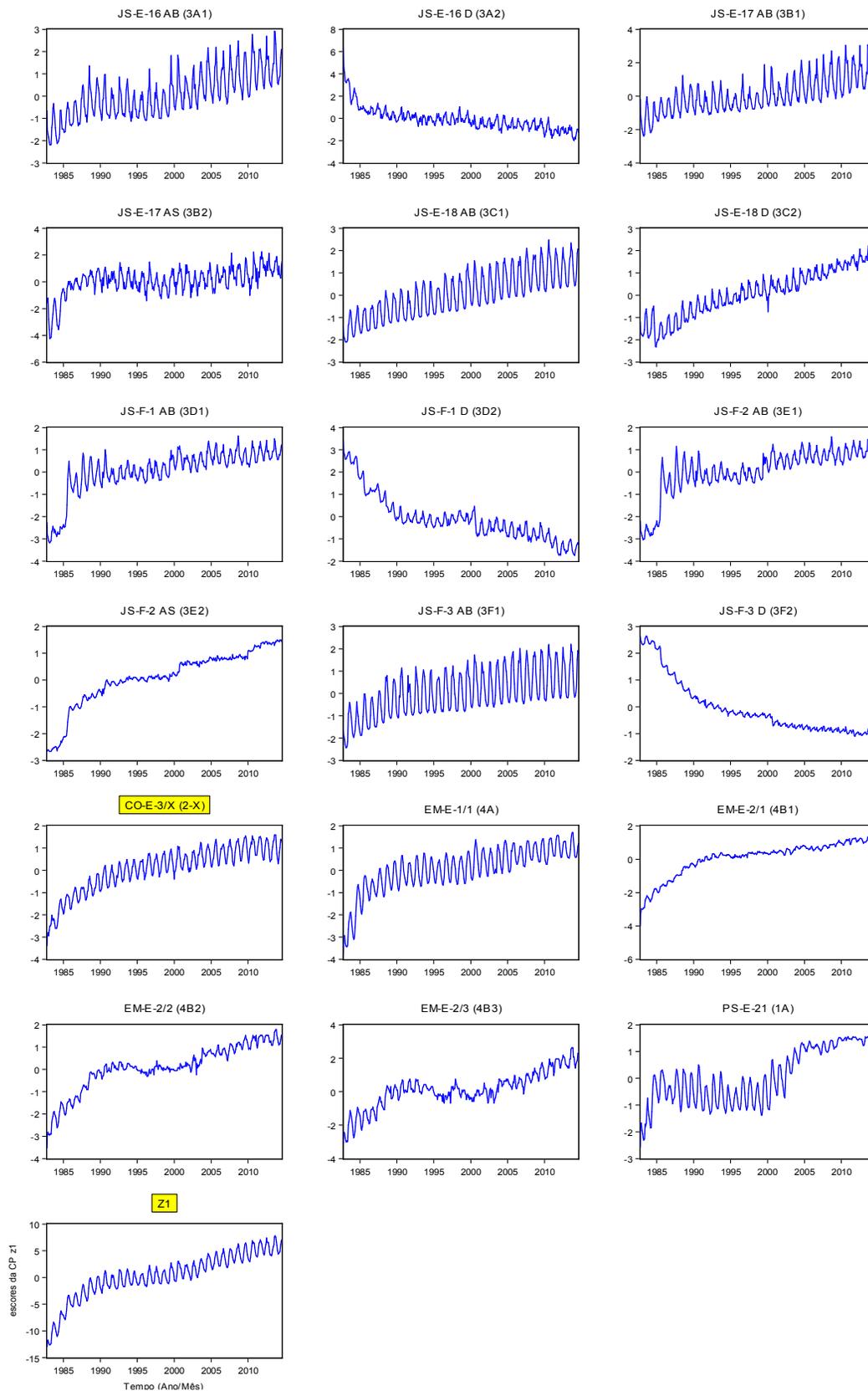


FIGURA 4.12 - Instrumentos mais importantes da CP z_1

FONTE: O autor, resultados da pesquisa

A CP z_2 é principalmente determinada pelos piezômetros 1B, 1C, 1F, 1G, 1H e o pêndulo direto na direção normal ao fluxo do rio z_2 . z_2 é 'aproximadamente' a soma destas variáveis, inclusive todos os coeficientes delas são positivos em z_2 , conforme a FIGURA 4.13 (acima). Os instrumentos mais importantes na determinação da CP z_3 são os extensômetros $4C_1$, $4C_2$ e $4C_3$. Nota-se na FIGURA 4.13 (abaixo) que a CP z_3 se aproxima da soma das contribuições do instrumento 4C nas hastes 1, 2 e 3. Estes instrumentos e a CP z_3 tem um comportamento muito similar. Importante citar que a componente conseguiu agrupar as três hastes do mesmo instrumento.

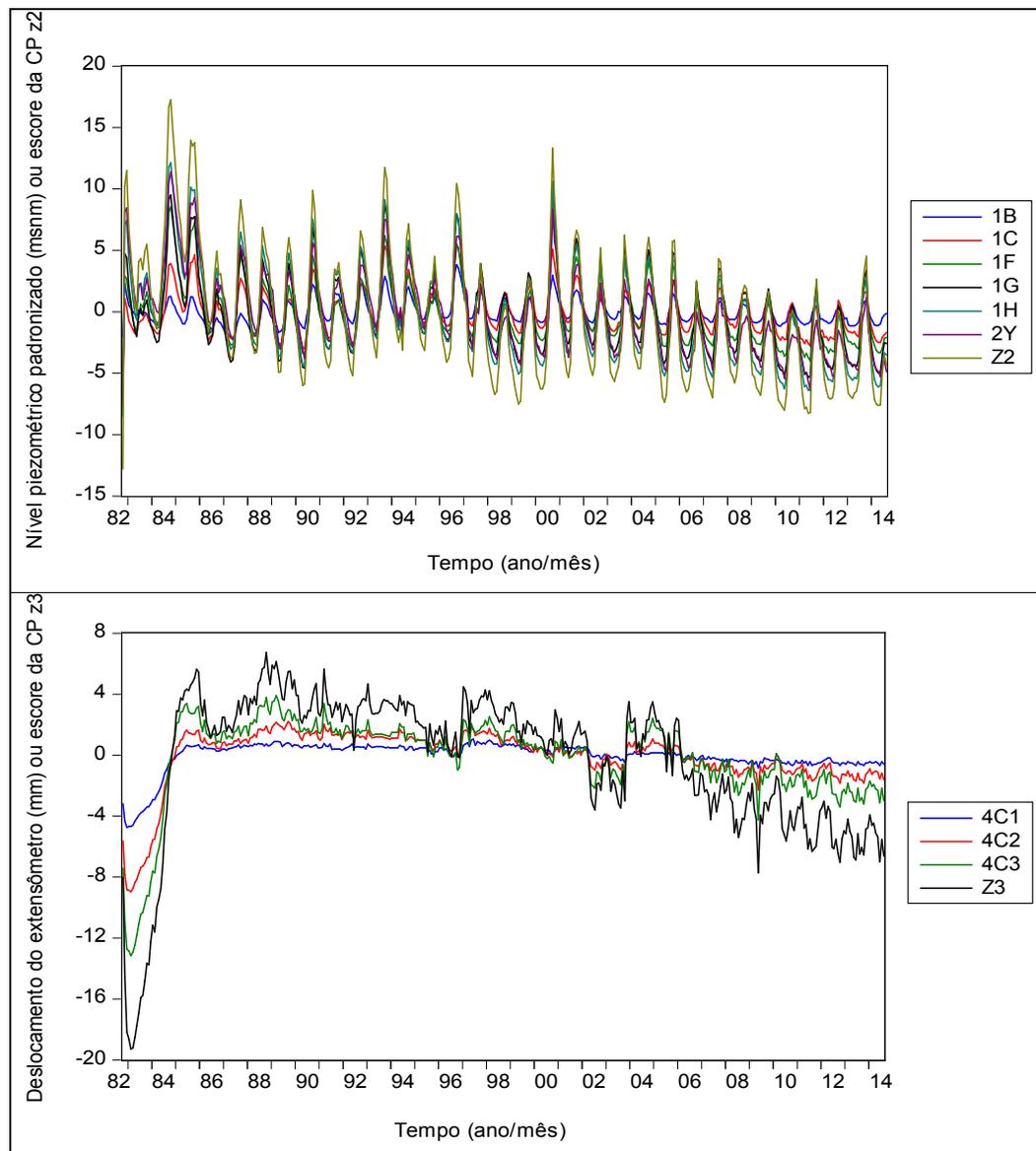


FIGURA 4.13 - Instrumentos mais importantes da CP z_2 (acima) e da CP z_3 (abaixo)

FONTE: O autor, resultados da pesquisa

A CP z_4 é principalmente determinada pelas três hastes do extensômetro 4D e o piezômetro 1D. Observe na FIGURA 4.14 (acima) que z_4 é praticamente a soma das contribuições das hastes 1,2,3 de 4D (positivamente) e de 1D (negativamente). A CP z_5 é determinada fundamentalmente pelo piezômetro 1E, conforme a FIGURA 4.14 (abaixo).

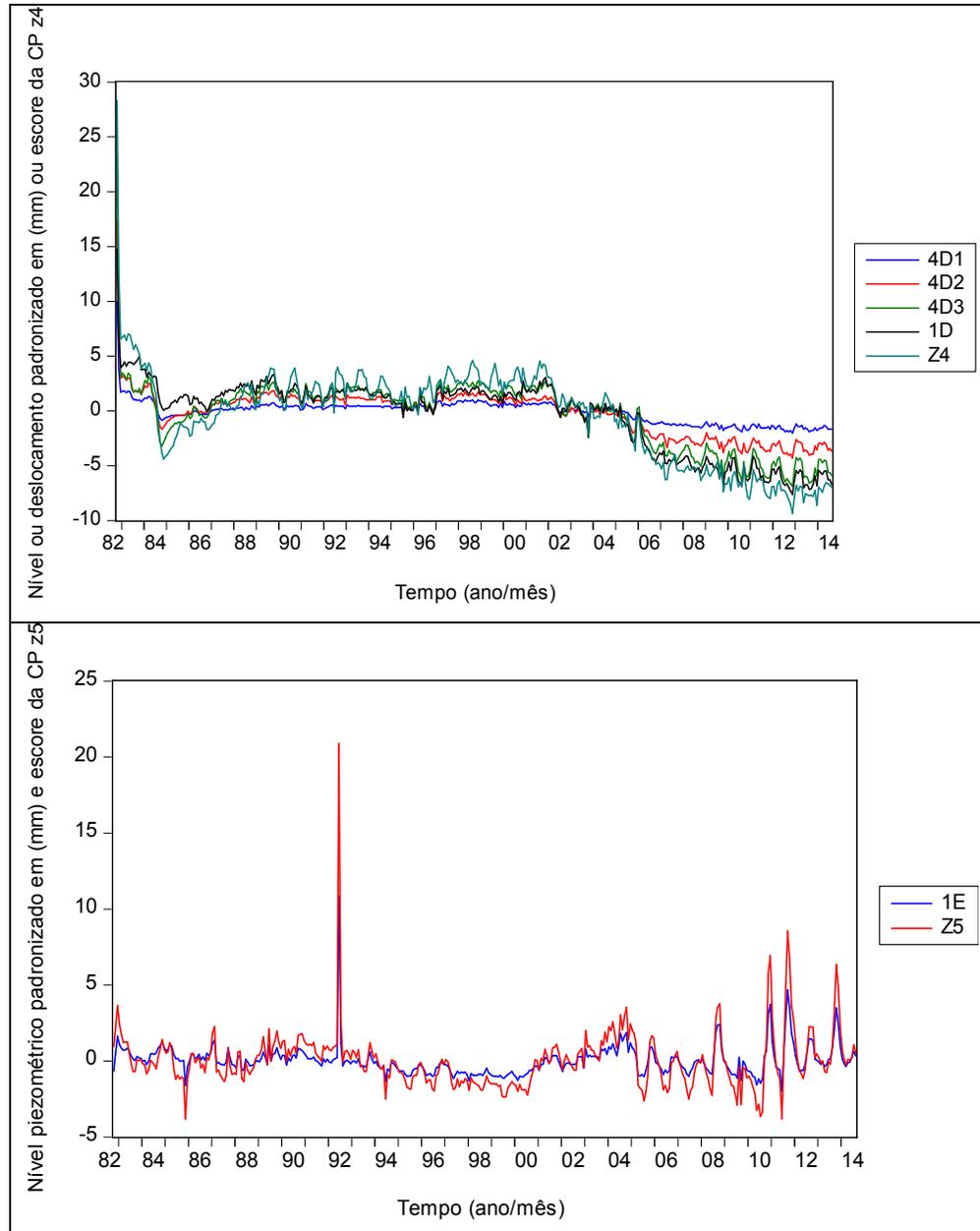


FIGURA 4.14 - Instrumentos mais importantes da CP z_4 (acima) e da CP z_5 (abaixo)

FONTE: O autor, resultados da pesquisa

A TABELA 4.23, apresenta os resultados do desempenho de previsão dinâmica do RMSE e do MAE nos períodos dentro (D), fora (F) e total (T) da amostra para 5 modelos adequados para z_1 .

TABELA 4.23 - Resultados da modelagem para CP z_1

MODELO	z_{1_1}			z_{1_2}			z_{1_3}			z_{1_4}			z_{1_5}		
	D	F	T	D	F	T	D	F	T	D	F	T	D	F	T
RMSE	0,29	10,38	10,75	0,63	3,34	3,55	0,67	3,27	3,48	0,68	2,77	3,03	0,83	3,06	3,29
MAE	0,25	9,58	9,94	0,50	3,09	3,26	0,54	3,00	3,18	0,56	2,34	2,55	0,67	2,75	2,93

FONTE: O autor, resultados da pesquisa

Dentre os modelos para z_1 , o modelo 4 apresenta melhor desempenho nos períodos fora e total e um pouco inferior no período dentro da amostra, por isso, resolveu-se selecionar o modelo 4, isto é, z_{1_4} . O modelo 4 é GARCH e tem menor valor do critério AIC (TABELA 4.22). Para as demais CP's z_2 a z_5 foi feita uma análise semelhante para a seleção dos melhores modelos. A TABELA 4.24 apresenta algumas características dos melhores modelos selecionadas para cada CP.

TABELA 4.24 - Resultados dos melhores modelos encontrados

CP	z_1	z_2	z_3	z_4	z_5
Melhor modelo	z_{1_4}	z_{2_2}	z_{3_1}	z_{4_1}	z_{5_2}
Resíduos	Res_{z_1}	Res_{z_2}	Res_{z_3}	Res_{z_4}	Res_{z_5}
R^2	99,38	96,03	94,62	94,54	59,06
Tempamb	0 a 10	0 a 3	4	0 a 3	Nenhum
DW	2,04	2,08	1,89	2,07	2,09
Arima	$ar(1), sar(12)$ $ma(1), sma(24)$	$ar(1), ar(3)$ $sar(12), ma(12)$	$ar(1)$	$Ar(1), ar(2), sar(12,)$ $ma(1), ma(7), ma(8)$	$Ar(1), ar(10),$ $ma(12)$
Garch(r, s)	(0,2)	--	--	--	--

FONTE: O autor, resultados da pesquisa

Note que, segundo a TABELA 4.25, todos os valores da probabilidade do teste de Jarque-Bera são menores do que 0,05, portanto, rejeita-se a normalidade univariada dos resíduos das CP's e como esta é condição necessária para a normalidade multivariada (LOONEY, 1995), deve se rejeitar a normalidade multivariada (MVN) dos resíduos das CP's e das CP's. A tentativa de transformação nos dados destas séries também não surtiu efeito para obter normalidade.

TABELA 4.25 - Teste de Jarque-Bera para normalidade dos resíduos das CP's e das CP's

Teste/Série	Res_z ₁	Res_z ₂	Res_z ₃	Res_z ₄	Res_z ₅	z ₁	z ₂	z ₃	z ₄	z ₅
Jarque – Bera	34.12	335.08	39.92	12.75	23409.17	62.31	16.49	43.90	3775.23	12551.42
Probabilidade	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

FONTE: O autor, resultados da pesquisa

Para a construção do gráfico T^2 dos resíduos, foram incluídas 338 observações na fase I (23 foram usadas no processo de estimativa dos parâmetros) e 361 na fase II. Valores do limite superior de controle (LSC) das fases I e II para taxas de falsos alarmes fixadas em $\alpha = 1\%, 2\%$ e 5% do gráfico T^2 e o número de observações fora dos limites de controle (FLC) para o gráfico T^2 dos resíduos e das CP's estão representadas na TABELA 4.26. Todos os processos podem ser declarados fora do limite de controle presumindo normalidade multivariada, pois a probabilidade de que n ou mais observações estejam acima do LSC_I para o α especificado, dado na quarta linha da TABELA 4.26, é sempre inferior a 0,05, portanto, rejeita-se a hipótese nula de que os dados estão sob controle ao nível de 95% de confiança.

TABELA 4.26 - LSC e quantidade de FLC para as fases I e II, $\alpha=1;2$ e 5% para os resíduos e para as CP's

α	Resíduos de z ₁ a z ₅			CPs de z ₁ a z ₅		
	1	2	5	1	2	5
LSC_I	14,85	13,22	10,97	14,87	13,23	10,97
FLC_I	12	17	28	18	21	32
Prob.	0,0002	0,0006	0,007	0,0000	0,0000	0,0013
LSC_{II}	15,09	13,38	11,07	15,08	13,38	11,07
FLC_{II}	12	18	29	19	20	33

FONTE: O autor, resultados da pesquisa

Para cada cenário da construção do KDE, se considerou os núcleos das equações (2.1.116) e (2.1.117), número de observações usadas no *kernel* $N = 200$ ou $N = 300$, h o parâmetro de suavização, equação (2.1.122), e foi obtido o percentil de 95, 98 e 99% que é o LSC exibido na TABELA 4.27 para a série temporal t^2_{res} bem como a quantidade de observações acima do LSC para a fase I (FLC I) e da fase II (FLC II). A TABELA 4.27 apresenta ainda o LSC e FLC presumindo MVN e usando

a melhor distribuição paramétrica plausível, que no caso foi uma log-logística. A TABELA 4.28 apresenta os mesmos resultados para a série temporal t^2_{pcs} .

TABELA 4.27 - Resultados dos valores do LSC e FLC obtidos via KDE para a série temporal t^2_{res}

Série		t^2_{res}				95%		98%		99%	
Núcleo	N	h	LSC 95%	LSC 98%	LSC 99%	FLC I	FLC II	FLC I	FLC II	FLC I	FLC II
N	200	0.78	13,17	16,46	23,70	19	20	8	8	5	5
N	200	0.39	13,70	16,95	24,11	17	18	8	8	5	5
N	300	0.78	13,33	17,25	25,98	17	18	8	8	5	5
N	300	0.39	13,16	16,60	26,07	19	20	8	8	5	5
E	200	1.74	13,18	16,46	23,04	19	20	8	8	5	5
E	200	0.87	13,05	16,30	22,81	19	20	9	9	6	6
E	300	1.74	13,33	16,82	25,55	17	18	8	8	5	5
E	300	0.87	13,24	16,27	25,81	17	18	9	9	5	5
MVN fase I			10,97	13,22	14,85	28		17		12	
MVN fase II			11,07	13,39	15,09		29		18		12
Log-logística			13,74	22,47	32,52	17	18	6	6	1	1
Esperado						16,9	18,1	6,76	7,22	3,38	3,61

FONTE: O autor, resultados da pesquisa

Nota-se que o LSC determinado via KDE ou distribuição log-logística faz com que o número de observações acima do limite superior de controle (FLC) para o gráfico T^2 esteja bem mais próximo da taxa de falsos alarmes esperada.

TABELA 4.28 - Resultados dos valores do LSC e FLC obtidos via KDE para a série temporal t^2_pcs

Série		t^2_pcs				95%		98%		99%	
Núcleo	N	h	LSC 95%	LSC 98%	LSC 99%	FLC I	FLC II	FLC I	FLC II	FLC I	FLC II
N	200	0.68	13,11	20,57	22,70	21	22	10	10	6	6
N	200	0.34	12,98	20,36	22,47	21	22	10	10	7	7
N	300	0.68	13,40	21,17	23,64	19	20	9	9	5	5
N	300	0.34	13,57	21,58	23,68	19	20	9	9	5	5
E	200	1.51	13,11	20,57	22,70	21	22	10	10	6	6
E	200	0.75	12,98	20,36	22,99	21	22	10	10	5	5
E	300	1.51	13,49	21,31	23,44	19	20	9	9	5	5
E	300	0.75	13,31	21,37	23,13	20	21	9	9	5	5
MVN fase I			10,98	13,23	14,87	32		21		18	
MVN fase II			11,07	13,39	15,09		33		20		19
Log-logística			12,2	19,2	27	22	23	10	10	3	3
Esperado						18	19,2	7,2	7,66	3,6	3,83

FONTE: O autor, resultados da pesquisa

Para exemplificar todos estes cenários, na

FIGURA 4.15 - Gráfico T^2 da série t^2_res , LSC MVN x KDE

é exibido o gráfico T^2 para a série t^2_res , $\alpha = 5\%$, incluindo dados das fases I e II, com o LSC determinado via normalidade multivariada (em vermelho; 11,07 conforme linha 10, coluna 4, TABELA 4.27, gerando 29 FLC) e determinado via KDE (em rosa; 13,33 conforme linha 3, coluna 4, TABELA 4.27, gerando 18 FLC) ou log-logística, sendo que o esperado é 18,1, conforme linha 12, TABELA 4.27.

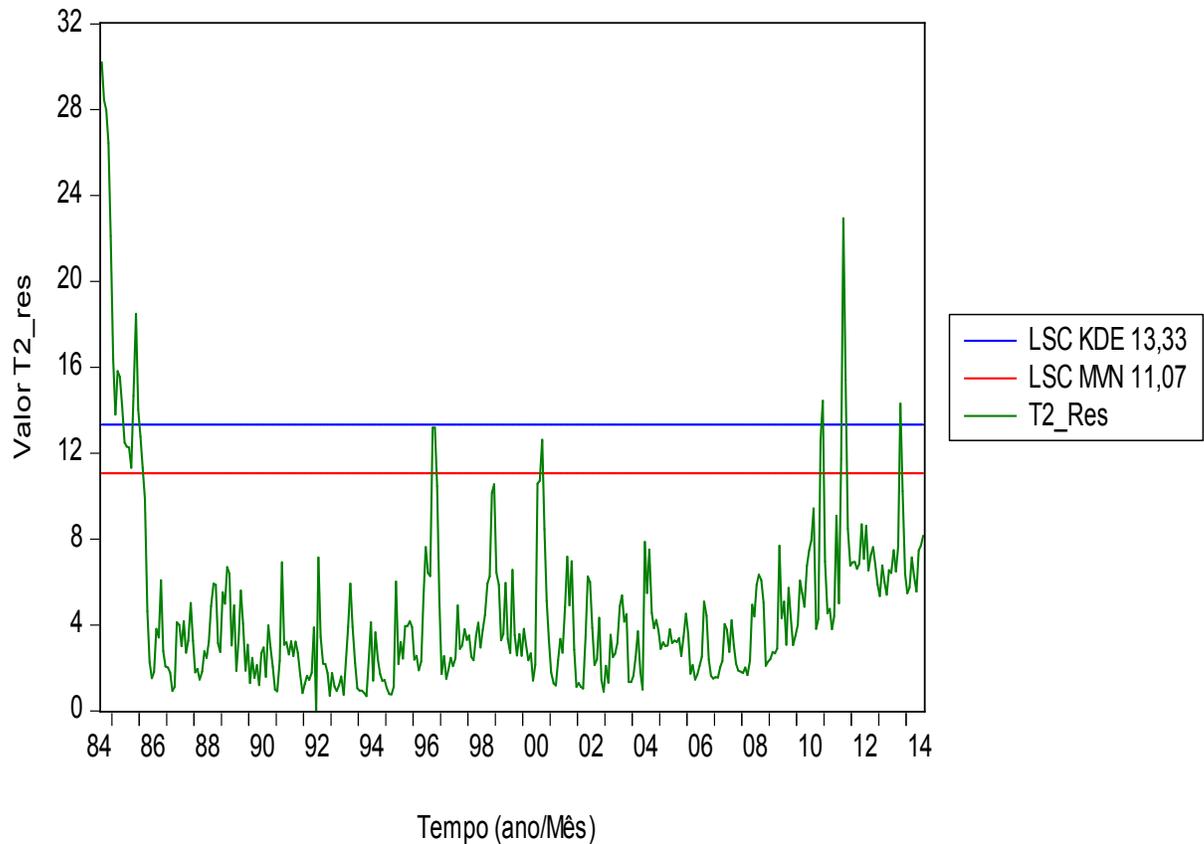


FIGURA 4.15 - Gráfico T^2 da série t^2_{res} , LSC MVN x KDE

FONTE: O autor, resultados da pesquisa

Isso mostra que, quando o LSC do gráfico T^2 é determinado sem presumir a normalidade multivariada, os falsos alarmes ficam dentro dos valores esperados e tornam viáveis os gráficos de controle multivariados aplicados ao monitoramento de instrumentos da barragem de Itaipu.

Esta etapa do trabalho multivariado que englobou quatro tipos de instrumentos e trinta e duas séries, mais a temperatura ambiente e o nível do lago, buscou determinar uma metodologia multivariada para monitorar dados de instrumentos de auscultação de barragens e estabelecer limites estatísticos de controle. Neste estudo de caso, buscou-se reduzir o modelo de monitoramento a algumas CP's representativas, monitorar o gráfico T^2 dos resíduos onde o limite superior de controle independe da hipótese de normalidade.

No experimento numérico, foram retidas cinco CP's, construídos modelos ARIMAX-GARCH e selecionado o melhor modelo em termos do MAE-RMSE para

ajuste dos dados e para previsões dinâmicas. Os resíduos destes modelos são monitorados com um gráfico T^2 e comparações da taxa de falsos alarmes são feitas com presunção da normalidade e de distribuições paramétricas plausíveis e não paramétricas via *KDE*.

A aplicação com dados da usina hidrelétrica de Itaipu (que pode ser estendida a outras barragens) mostrou que o procedimento proposto consegue reduzir a análise de trinta e dois instrumentos à análise de cinco componentes principais não correlacionadas e o LSC determinado via KDE ou distribuição log-logística reduz a taxa de falsos alarmes e dá maior credibilidade ao sistema de monitoramento. Os melhores cenários, comparativamente ao número esperado de falsos alarmes, são quando se usam 300 pontos na malha do KDE e o valor padrão do parâmetro h . A retirada das causas de variação atribuíveis (temperatura ambiente significativa) nos modelos ARIMAX-GARCH também foi fundamental uma vez que o número de observações fora dos limites de controle é menor nos resíduos.

O modelo de CP's conseguiu fazer um agrupamento do conjunto de instrumentos em cinco conjuntos disjuntos. Como a hipótese de MVN não é satisfeita, é gerada uma quantidade inaceitável de falsos alarmes no gráfico T^2 . Esta restrição do problema foi superada pelo uso do LSC obtido via KDE ou a distribuição log-logística.

A temperatura ambiente e seus *lags* de retardo comprovaram ser uma variável importante nos modelos ARIMAX-GARCH das CP's e, por consequência, das variáveis originais, os instrumentos. A variável do nível do reservatório não se mostrou significativa nos modelos utilizados. A justificativa para isso pode ser a baixa variabilidade do nível do reservatório em Itaipu.

Uma interpretação física para as componentes principais, discutida com um engenheiro da equipe segurança da barragem, é que a componente 1 desta análise representa o deslocamento do bloco (trecho) na direção do fluxo do rio; a componente 2 o deslocamento do bloco (trecho) na direção normal ao fluxo; a componente 3 avalia o recalque no bloco; a componente 4 a compressão a jusante do bloco e a componente 5 demonstra que o piezômetro 5 (PS-E-26), que tem um comportamento diferente dos demais, com pouca variabilidade, apresenta-se seco

e por situar-se após um túnel de drenagem no bloco, acaba por confirmar a eficiência do túnel de drenagem.

5 CONCLUSÕES

Este trabalho, de uma forma geral, teve o objetivo de contribuir no aspecto da aplicação de métodos numéricos em um problema de engenharia, da determinação de gráficos de controle univariados e multivariados para séries temporais de dados de leituras de instrumentos de monitoramento de barragens e, também, do ponto de vista acadêmico.

Do ponto de vista da aplicação tornou-se possível:

- Estabelecer valores estatísticos de controle para gráficos univariados de leituras de instrumentos de monitoramento num trecho da barragem de Itaipu e atualizar os estudos de Itaipu em relação a este tema;
- Propor a utilização de modelos estatísticos de gráficos de controle multivariados ainda não empregados em Itaipu;
- Auxiliar na identificação e eventual remoção de observações discrepantes (*outliers*) nos gráficos de controle de instrumentação de Itaipu;
- Colaborar com mais uma ferramenta que reitera a cultura da segurança de barragem em Itaipu.

Do ponto de vista acadêmico tornou-se possível:

- Promover, por meio das combinações híbridas, uma melhoria no desempenho da previsão de séries temporais univariadas e multivariadas de dados em Itaipu;
- Construir gráficos de controle para dados multivariados autocorrelacionados e correlacionados que não seguem a distribuição normal multivariada com menor taxa de falsos alarmes comparativamente aos gráficos tradicionais.

6 TRABALHOS FUTUROS

Durante a realização de uma tese, sempre existem fatores que limitam o trabalho, seja este o tempo para absorver novos conhecimentos e realizar simulações computacionais ou a mera necessidade de fixar objetivos tangíveis. Desta forma, são citadas a seguir algumas sugestões de assuntos ou ideias que não puderam ser abordados neste trabalho e que talvez possam colaborar no estabelecimento de limites estatísticos de controle e previsões mais precisas para dados de instrumentos de barragens.

Sugere-se o estudo de componentes principais não lineares, que podem agregar a estrutura de dependência não linear existente entre as componentes principais, abordar o erro tipo II, ou seja, o falso positivo, aquela observação que é dada como estando sob controle, quando na verdade não está. Esta é uma tarefa das mais importantes, mas depende inicialmente da definição de uma 'boa' região de controle, onde ocorre o erro tipo I.

Também, podem ser trabalhadas outras bases da decomposição Wavelet, além da base de Haar, bem como, outros níveis de decomposição, além disso, a automatização na programação de algumas Redes Neurais pode tornar as previsões mais acuradas e precisas.

A implementação de variáveis exógenas (temperatura ambiente, nível do reservatório) para explicar as componentes principais é o ponto de partida para trabalhar com modelos de duplo bloco, com múltiplas variáveis de entrada e saída como mínimos quadrados parciais e não se descarta a análise de correlação canônica.

As metodologias aplicadas a dados da barragem de Itaipu podem (e devem) ser avaliadas em outros dados de monitoramento de barragens e outras bases de dados utilizadas na literatura.

REFERÊNCIAS

ADHIKARI, R.; AGRAWAL, R. K. **An introductory study on time series modelling and forecasting**. [S.l.]: Lambert Academic Publishing, 2013.

ASDSO - ASSOCIATION OF STATE DAM SAFETY OFFICIALS, 2015. Disponível em: <<http://www.damsafety.org/resources/?p=362e7deb-1d29-4659-ace7-1f6b28eea34d>>. Acesso em: 5 novembro 2015.

BERSIMIS, S.; PANARETOS, J.; PSARAKIS, S. Multivariate statistical process control charts and the proble of interpretation: a short overview and some applications in industry. **Proceedings of the 7th. Hellenic European Conference on Computer Mathematichs ant its applications**, Atenas, p. 1-6, 2005.

BERSIMIS, S.; PSARAKIS, S.; PANARETOS, J. Multivariate statistical process control charts: an overview. **Quality and reliability engineering international**, 2007. 517-543.

BIASE, A. G.; FERREIRA, D. F. Teste computacionalmente intensivo baseado na distância de Mahalanobis para normalidade multivariada. **Rev. Bras. Biom.**, São Paulo, v. 30, n. 1, p. 1-22, 2012.

BOLLERSLEV, T. Generalized Autorregresive Conditional Heteroscedasticity. **Journal of Econometrics**, v. 31, n. 3, p. 307-326, Apr. 1986. doi:10.1016/0304-4076(86)90063-1.

BOONE, J. M. **Contributions to multivariate control charting**: studies of the Z-Chart and four nonparametric charts. University of Alabama: Doctoral Dissertation, 2010.

BOONE, J. M.; CHAKRABORTI, S. Two simple Shewhart-type multivariate nonparametric control charts. **Applied Stochastic Models in Business and Industry**, v. 28, p. 130-140, 2012. doi: 10.1002/asmb.900.

BOX, G. E. P.; JENKINS, G. W.; REINSEL, G. C. **Time Series Analysis Forecasting and Control**. 4a. ed. Hoboken-NJ: John Wiley & Sons, Inc., 2008.

BOX, G. E. P.; TIAO, G. C. Intervention Analysis with Applications to Economic and Environmental Problems. **Journal of the American statistical association**, v. 70, n. 349, p. 70-79, march 1975.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais: Teoria e aplicações**. Rio de Janeiro : LTC, 2000.

BUZZI, M. F. **Avaliação das correlações de séries temporais de leituras de instrumentos de monitoração geotécnico estrutural e variáveis ambientais em**

barragens - estudo de caso de Itaipu. Curitiba: Dissertação de Mestrado - PPGMNE - UFPR, 2007.

CAO, M.; QIAO, P.; REN, Q. Improved hybrid wavelet neural network methodology for time-varying behavior prediction of engineering structures. **Neural Comput. & Applic.**, v. 18, p. 821-832, 2009.

CASA CIVIL DA PRESIDÊNCIA DA REPÚBLICA. Lei nº 12.334. **Site da Casa Civil**, Brasília, 20 setembro 2010. Disponível em: <www.planalto.gov.br/ccivil_03/_Ato2007-2010/2010/Lei/L12334.htm>. Acesso em: 07 abril 2014.

CHAKRABORTI, S. Nonparametric (distribution-free) quality control charts. **Encyclopedia of Statistical Sciences - John Wiley & Sons, Inc.**, p. 1-27, 15 August 2011. DOI: 10.1002/0471667196.ess7150.

CHAKRABORTI, S.; VAN DER LAAN, P.; BAKIR, S. T. Nonparametric control charts: an overview and some results. **Journal of quality technology**, v. 33, p. 304-315, 2001.

CHATFIELD, C. **The analysis of time series**. 5a. ed. London: Chapman & Hall/CRC, 1996.

CHAVES NETO, A. **Bootstrap em séries temporais**. Rio de Janeiro: Tese de doutorado, PUC-RJ, 1991.

CHENG, L.; ZHENG, D. Two online dam safety monitoring models based on the process of extracting environmental effect. **Advances in Engineering Software**, v. 57, p. 48-56, March 2013. doi:10.1016/j.advengsoft.2012.11.015.

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

CIGB-ICOLD. **80 Years - Dams for human sustainable development**. Paris: International Commission on Large Dams , 2008.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. New York: Cambridge University Press, 1997.

DÍAZ-ROBLES, L. A. *et al.* A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas; The case of Temuco, Chile. **Atmospheric Environment**, v. 42, p. 8331-8340, 2008.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. Boca Raton-FL: Chapman & Hall/CRC, 1993.

ENGLE, R. F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom inflation. **Econometrica**, v. 50, n. 4, p. 987-1007, July 1982. <http://www.jstor.org/stable/1912773>.

FERNANDES, C. **Análise de séries temporais**. Rio de Janeiro: Laboratório de Estatística Computacional - PUC-RJ, 2006.

FOUCART, T. A decision rule for discarding principal components in regression. **Journal of Statistical Planning and Inference**, v. 89, n. 1-2, p. 187-195, August 2000.

FUSARO, 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.

G1. **Globo**, 2014. Disponível em: <www.g1.globo.com/pr/oeste-sudoeste/noticia/2014/05/seria-inimaginavel-o-paraguai-e-o-brasil-sem-itaipu-diz-diretor-geral.html>. Acesso em: 08 maio 2014.

GAZETA DO POVO. Gazeta do Povo, 08 Maio 2014. Disponível em: <www.gazetadopovo.com.br/economia/especial-itaipu-30-anos/conteudo.phtml?tl=1&id=1466225&tit=Gigante-pela-propria-natureza>. Acesso em: 08 Maio 2014.

GOLDBARG, M. C.; LUNA, H. P. L. **Otimização combinatória e programação linear**. 7a. ed. Rio de Janeiro: Elsevier, 2000.

GU, C. *et al.* Singular value diagnosis in dam safety monitoring effect values. **Science China Technological Sciences**, v. 54, n. 5, p. 1169-1176, May 2011. doi: 10.1007/s11431-011-4339-7.

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

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton University Press, 1994.

HINES, W. W. *et al.* **Probabilidade e estatística na engenharia**, 4a. ed. Rio de Janeiro: LTC, 2006.

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

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

HOTELLING, H. Multivariate quality control. **Techniques of statistical analysis**, p. 111-184, 1947.

ICOLD - BULLETIN 21. **General considerations applicable to instrumentation for earth and rockfill dams**. Paris. 1969.

ICOLD - CIGB. **Dams & The World's Water - An educational book that explains how dams help to manage the world's water**. Paris: International Commission on Large Dams/Commission Internationale des Grand Barrages, 2008.

ITAIPU BINACIONAL. **Síntese das Principais Informações de Projeto para o Controle da Instrumentação da Barragem Principal - Trechos E, F e I (4106.50.06001)**. Foz do Iguaçu. 1984.

ITAIPU BINACIONAL. **Aspectos teóricos dos Estudos realizados - Estudos para a reavaliação dos valores de controle para a instrumentação civil das estruturas e fundações (2660.50.15402.P)**. Foz do Iguaçu. 1999_a.

ITAIPU BINACIONAL. **Bloco de Contrafortes Bloco E-6 - Estudos para a reavaliação dos valores limites para a instrumentação civil das estruturas e fundações (2662.50.15401-P)**. Foz do Iguaçu. 1999_b.

ITAIPU BINACIONAL. **Modelo Estatístico - Estudos para a reavaliação dos valores de controle para a instrumentação civil das estruturas e fundações (2660.50.15401 P)**. Foz do Iguaçu. 1999_c.

ITAIPU BINACIONAL. **Valores de Controle para os instrumentos integrantes do "ADAS" e critérios para a geração de alarmes**. Foz do Iguaçu. 2002.

JACKSON, J. E. Principal components and factor analysis: part I - Principal components. **Journal of quality technology**, v. 12, n.4, p. 201-213, 1980.

JACKSON, J. E. Multivariate quality control. **Communications in statistics - Theory and Methods**, v. 14, n. 11, p. 2657-2688, 1985. DOI:10.1080/03610928508829069.

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

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

JEON, J. *et al.* Development of dam safety management system. **Advances in Engineering Software**, p. 554-563, 2009.

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

JOLLIFFE, I. T. **Principal Component Analysis, 2nd edition**. New York: Springer, 2002.

JORNAL INTERNO DE ITAIPU - JIE. **Jornal de Itaipu Eletrônico**, 08 maio 2014. Disponível em:

<jie.itaipu.gov.br/index.php?secao=noticias_itaipu&q=node/6&pagina=4>. Acesso em: 08 maio 2014.

KALGONDA, A. A.; KULKARNI, S. R. Multivariate quality control chart for autocorrelated process. **Journal of applied statistics**, v.31, n. 3, p. 317-327, 2004.

KHASHEY, M.; BIJARI, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. **Applied soft computing**, v. 11, p. 2664-2675, 2011.

KHEIFITS, V. Z. Routine instrumental monitoring - A guarantee of dam safety. **Hydrotechnical Construction**, v. 29, n.5, p. 287-289, 1995.

KOURTI, T. Application of latent variable methods to process control and multivariate statistical process control in industry. **International Journal of adaptive control and signal processing**, v.19, p. 213-246, 2005.

KRIECHBAUMER, T. *et al.* An improved wavelet-ARIMA approach for forecasting metal prices. **Resources Policy**, v. 39, p. 32-41, 2014.

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

KUBRUSLY, C. S. **The Elements of Operator Theory**. 2nd. ed. Boston: Birkhauser, 2011.

KUBRUSLY, C. S.; LEVAN, N. Abstract Wavelets Generated by Hilbert Space Shift Operators. **Adv. Math. Sci. Appl.**, v. 16, p. 643-660, 2006.

KUMAR, U.; DE RIDDER, K. GARCH modelling in association with FFT e ARIMA to forecast ozone episodes. **Atmospheric Environment**, v. 44, p. 4256-4265, 2010.

KUPERMAN, S. C. *et al.* **Criteria to establish limit values of instrumentation readings for old embankment and concrete dams**. ASDSO Annual Conference. [S.I.]: [s.n.]. 2004.

KUPERMAN, S. C. *et al.* Placing 'limit values' on instruments readings. **HydroWorld**, v. 15, n. 3, p. 24-31, July 2007.

LAZZAROTTO, E. *et al.* A Wavelet SARIMA-ANN linear combination with multiple stages in time series prediction. **International Journal of Engineering and Applied Sciences**, v. 7, n. 4, p. 8-16, November 2015. ISSN 2305-8269. Disponível em: <<http://eaas-journal.org/survey/userfiles/files/v7i402%20Artificial%20Neural%20Networks.pdf>>.

LAZZAROTTO, E. *et al.* **Combinação de análise de componentes principais com técnicas não paramétricas para estabelecer valores limite em gráficos de**

controle aplicados em dados de instrumentação de barragem. CMN 2015 - Congress on Numerical Methods in Engineering. Lisboa: APMTAC-SEMNI. 2015.

LAZZAROTTO, E. *et al.* Melhoria na previsão da série temporal de instrumentos de monitoramento de barragem via combinação de métodos. **Cadernos do IME - Série Estatística**, v. 38, Junho 2015. ISSN 1413-9022. [dx.doi.org/10.12957/cadest](https://doi.org/10.12957/cadest).

LAZZAROTTO, E. *et al.* Principal Components in multivariate control charts applied to data of instrumentation of dams. **Independent Journal of Management & Production**, v. 7, n. 1, p. 17-37, January - March 2016. ISSN 2236-269X. DOI: 10.14807/ijmp.v7i1.369.

LAZZAROTTO, E.; GRAMANI, L. M.; CHAVES NETO, A. Comparação de gráficos de controle univariados e multivariados aplicados a dados de instrumentação da usina hidrelétrica de Itaipu. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics - XXXV CNMAC - Congresso Nacional de Matemática Aplicada e Computacional**, Natal-RN, v. 3, n. 1, setembro 2014. Disponível em: <<http://proceedings.sbmac.org.br/sbmac/issue/view/3/showToc>>. DOI: 10.5540/03.2015.003.01.0266.

LAZZAROTTO, E.; GRAMANI, L. M.; CHAVES NETO, A. Estabelecimento de valores de controle em um gráfico de controle univariado autocorrelacionado. **IV Simpósio de Métodos Numéricos Computacionais da Universidade Federal do Paraná**, Curitiba, novembro 2014. ISSN 2236-8108, v.1, n.4.

LAZZAROTTO, E.; GRAMANI, L. M.; CHAVES NETO, A. Técnica não paramétrica para estabelecer valores limite em gráficos de controle aplicados em dados de instrumentação de barragem. **Comitê Brasileiro de Grandes Barragens - XXX Seminário Nacional de Grandes Barragens**, Foz do Iguaçu, maio 2015.

LIANG, J. Multivariate statistical process monitoring using kernel density estimation. **Dev. Chem. Eng. Mineral Process**, v. 13, n. 1-2, p. 185-192, 2005. DOI: 10.1002/apj.5500130117.

LIU, H.; SHI, J. Applying ARIMA-GARCH approaches to forecasting short-term electricity prices. **Energy Economics**, v. 37, p. 152-166, 2013.

LOH, C.; CHEN, C. H.; HSU, T. Y. Application of advanced statistical methods for extracting long-term trend in static monitoring data from an arch dam. **Structural Health Monitoring**, v. 10, p. 587-601, 11 January 2011. doi: 10.1177/1475921710395807.

LOONEY, S. W. How to use tests for univariate normality to assess multivariate normality. **The American Statistician**, v. 49, n. 1, p. 64-70, Feb. 1995. DOI: 10.2307/2684816.

LOWRY, C. A.; MONTGOMERY, D. C. A review of multivariate control charts. **IIE Transactions**, v. **26**, 1995.

LUTKEPOHL, H. **New Introduction to multiple time series analysis**. [S.l.]: Springer, 2006.

MALLAT, S. **A wavelet tour of signal processing**. 2nd. ed. San Diego: Academic Press, 1998.

MARTIN, E. B.; MORRIS, A. J. Non-parametric confidence bounds for process performance monitoring charts. **J. Proc. Cont.**, v. **6**, n. **6**, p. 349-358, 1996.

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

MASTRANGELO, C. M.; RUNGER, G. C.; MONTGOMERY, D. C. Statistical process monitoring with principal components. **Quality and reability engineering international**, v. **12**, p. 203-210, 1996.

MATA, J.; CASTRO, A. T.; COSTA, J. S. Constructing statistical models for arch dam deformation. **Structural Control and Health Monitoring**, v. **21**, p. 423-437, 26 april 2014. DOI: 10.1002/stc.1575.

MATHWORKS. MatLab - The language of technical computing, 2015. Disponivel em: <<http://www.mathworks.com/products/matlab/>>. Acesso em: 12 março 2015.

MATHWORKS. MatLab - The language of technical computing, 2016. Disponivel em: <<http://www.mathworks.com/help/wavelet/>>. Acesso em: 14 Março 2016.

MILETIC, I. *et al.* Review - An industrial perspective on implementing on-line applications of multivariate statistics. **Journal of Process Control**, v. **14**, p. 821-836, 2004.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: UFMG, 2005.

MINISTÉRIO DAS MINAS E ENERGIA. Secretaria de Planejamento e Desenvolvimento Energético. **Boletins Atuais**, 23 Maio 2013. Disponivel em: <www.mme.gov.br/mme/galerias/arquivos/publicacoes/boletins_de_energia/boletins_atuais/22_-_Potxncia_Instalada_EE.pdf>. Acesso em: 07 Abril 2014.

MINISTÉRIO DE MINAS E ENERGIA. Secretaria de Energia Elétrica - Departamento de Monitoramento do Sistema Elétrico. **Boletim Mensal de Monitoramento do Sistema Elétrico Brasileiro - Janeiro/2016**, 2016. Disponivel em: <http://www.mme.gov.br/web/guest/pagina-inicial/outras-noticias/-/asset_publisher/32hLrOzMkWWb/content/capacidade-instalada-de-geracao-de-energia-eletrica-atinge-141-684-mw-em-janeiro>. Acesso em: 16 Março 2016.

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

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

MURTY, K. G. **Linear Programming**. New York: John Wiley & Sons, 1983.

MYERS, B. K.; MARILEY, J. M. Automated monitoring at Tolt dam. **Civil Engineering**, v.67, n. 3, p. 44-46, 1997.

NATIONAL INSTITUTE OF STANDARDS TECHNOLOGY. Information Technology Laboratory. **Engineering statistics handbook**, 2014. Disponível em: <<http://www.itl.nist.gov/div898/handbook/pmc/section1/pmc12.htm>>. Acesso em: 21 agosto 2014.

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

NIAKI, S. T. A.; ABBASI, B. Bootstrap method approach in designing multi-attribute control charts. **Int. J. Adv. Manuf. Technol.** v.35, p. 434-442, 2007.

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

OPERADOR NACIONAL DO SISTEMA ELÉTRICO. **ONS - Avaliação de Curto Prazo da Operação - Séries Históricas de Vazões**, 2016. Disponível em: <www.ons.org.br/operacao/vazoes_naturais.aspx>. Acesso em: 20 julho 2015.

PAN, X.; JARRETT, J. E. Why and how to use vector autoregressive models for quality control: the guideline and procedures. **Quality and quantity**, v. 46, p. 935-948, 2012.

PANKRATZ, A. **Forecasting with dynamic regression**. 1. ed. [S.l.]: Wiley-Interscience, 1991.

PHALADIGANON, P. *et al.* Bootstrap-based T2 Multivariate control charts. **Communications in Statistics - Simulation and Computation**, v. 40, p. 645-662, 2011.

PHALADIGANON, P. *et al.* Principal component analysis-based control charts for multivariate nonnormal distributions. **Expert Systems with Applications**, v. 40, n. 8, p. 3044-3054, 15 June 2013. doi:10.1016/j.eswa.2012.12.020.

PINTO, L. P. **Estudo comparativo de testes de hipótese multivariados para matrizes de covariâncias via simulação de Monte Carlo e suas aplicações em controle de qualidade**. Belo Horizonte: Dissertação de Mestrado - UFMG, 2009.

PINTO, W. J.; SILVA, O. M. **Econometric Views - Guia do Usuário**, 2015. Disponível em: <<http://www.ufv.br/dee/ApostilaEviews.pdf>>. Acesso em: 12 março 2015.

QUINTANILHA, J. A. *et al.* Use of control charts for multi-temporal analysis of geodetic auscultation data from dams. **BCG - Boletim de Ciências Geodésicas**, Curitiba, v. 19, n. 4, p. 653-666, out-dez 2013. <http://dx.doi.org/10.1590/S1982-21702013000400009>.

RAGSDALE, C. F. **Spreadsheet Modeling & Decision Analysis - A practical introduction to management science**. 6. ed. Mason-OH: South-Western Cengage Learning, 2012.

ROSSO, J. A. *et al.* **Barragem de Itaipu - Lições aprendidas com o comportamento térmico das estruturas tipo gravidade aliviada e contraforte**. Foz do Iguaçu. 1995.

ROYSTON, J. P. Some techniques for assessing multivariate normality on the Shapiro-Wilk W. **Journal of the Royal Statistical Society. Series C (Applied Statistic)**, v. 32, n. 2, p. 121-133, 1983.

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.

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

SALAZAR, F. *et al.* Data-Based Models for the Prediction of Dam Behaviour: A Review and Some Methodological Considerations. **Archives of Computational Methods in Engineering**, p. 1134-1160, 25 July 2015. DOI 10.1007/s11831-015-9157-9.

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

SEBER, G. A. F. **Multivariate observations**. Hoboken-NJ: John Wiley & Sons., 2004.

SILVEIRA, J. F. A. **Instrumentação e comportamento de fundações de barragens de concreto**. São Paulo: Oficina de Textos, 2003.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. London: Springer Science+Business Media B.V., 1986.

STATGRAPHICS Centurion XVI. Disponível em: <<http://www.statgraphics.com/>>. Acesso em: 28 abril 2014.

SUN, R.; TSUNG, F. A kernel-distance-based multivariate control chart using support vector methods. **International Journal of Production Research**, v.41, n. 13, p. 2975-2989, 2003.

TAN, Z. *et al.* Day ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. **Applied energy**, v. 87, p. 3606-3610, 2010.

TAWARMALANI, M.; SAHINIDIS, N. **Convexification and global optimization in continuous and mixed-integer nonlinear programming - theory algorithms, software, and applications**. [S.l.]: Springer, 2002.

TEIXEIRA JR., L. A. **Combinação SSA-wavelet de métodos preditivos com ajuste numérico MIMIMAX, na geração de previsões e cenários**. Rio de Janeiro: Tese de doutorado - Departamento de Engenharia Elétrica - PUC, 2013.

UOL. Economia UOL. **UOL**, 2016. Disponível em: <<http://economia.uol.com.br/noticias/redacao/2016/01/07/itaipu-bate-chinesa-e-volto-a-ser-lider-em-producao-de-energia-em-2015.htm>>. Acesso em: 07 janeiro 2016.

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

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.

VILLWOCK, R. **Técnicas de agrupamento e de hierarquização no contexto de KDD – Aplicação a dados temporais de instrumentação geotécnico-estrutural da usina hidrelétrica de Itaipu**. Curitiba: Tese de Doutorado - UFPR - PPGMNE, 2009.

VOGELVANG, B. **Econometrics - Theory and application with Eviews**. Harlow-England: Pearson Education Limited, 2005.

WALNUT, D. F. **An introduction to wavelet analysis**. Boston: Birkhauser, 2002.

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

WOODALL, W. *et al.* Using control charts to monitor process and product quality profiles. **Journal of quality technology**, v.36, n.3, p. 309-320, 2004.

XU, B. S. *et al.* Analysis method of thermal dam deformation. **Science China Technological Sciences**, v. 55, n. 6, p. 1765-1772, June 2012. DOI 10.1007/s11431-012-4839-0.

XXIII ICOLD QUESTION 88. **Hydropower current developments**. Dams and Hydropower. Brasília: ICOLD - International Comitee of Large Dams. 2009.

XXIII ICOLD QUESTION 90. **Upgrading of existing dam**. Dams and Hydropower. Brasília: ICOLD - International Comitee of Large Dams. 2009.

XXIII ICOLD QUESTION 91. **Dam safety management**. Dams and Hydropower. Brasília: ICOLD - International Comitee of Large Dams. 2009.

XXIV ICOLD QUESTION 93. **Dam safety incidents**. Established data bases and needs. Kyoto: ICOLD - International Comitee of Large Dams. 2012.

YANG, L.; SHEU, S. Integrating multivariate engineering process control and multivariate statistical process control. **International Journal Adv Manuf Technol**, v.29, p. 129-136, 2006.

YANG, X.-S. **Introduction to Mathematical Optimization – From Linear Programming to Metaheuristics**. Cambridge: Cambridge International Science Publishing, 2008.

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

ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, v. 50, p. 159-175, 2003.

APÊNDICE A (CNMAC)

XXXV Congresso Nacional de Matemática Aplicada e Computacional

Os instrumentos de monitoramento de barragens em usina hidrelétricas fornecem extensas bases de dados das quais podem ser retiradas importantes informações. Uma técnica adequada para tal são os gráficos de controle estabelecidos com base em critérios estatísticos. Neste trabalho é realizada a interpretação e a comparação entre gráficos de controle básicos, os gráficos univariados \bar{x} e os gráficos bivariados T^2 de Hotelling aplicados aos dados de leitura de piezômetros localizados no trecho E da barragem da usina hidrelétrica de Itaipu. Resultados demonstram que os gráficos multivariados são capazes de reduzir a taxa de falsos alarmes, facilitarem a identificação de *outliers*, apontar uma das principais causas de variabilidade nestes gráficos de controle e estabelecer limites de controle operacionais para as leituras de instrumentos da usina. (LAZZAROTTO, GRAMANI e CHAVES NETO, 2014)

APÊNDICE B (SMNC)

IV Simpósio de Métodos Numéricos Computacionais

Este trabalho buscou eliminar o efeito da autocorrelação serial existente na série temporal da leitura de um instrumento de monitoramento de barragem. Sem o uso de modelos ARIMA, de um total de 300 dados, mais da metade se apresentava fora do limite de controle do gráfico de controle de Shewhart da média. O uso de modelos ARIMA, combinado com testes subsequentes de normalidade dos resíduos, permitiu inicialmente reduzir o número de dados fora dos limites de controle para 5. Investigações adicionais levaram a identificação de 4 *outliers* nestes dados fora de controle e que sua remoção permite o estabelecimento de limites de controle, a um nível de confiança de 95%, para um conjunto histórico de dados. No caso da taxa de falsos alarmes estabelecida, para uma amostra de 300 dados, isto requer que nenhum valor fique dos limites de controle. A combinação dos resíduos de modelos ARIMA com o gráfico de controle da média amostral, aplicada ao caso de dados de leitura de um instrumento de monitoramento em Itaipu, permitiu a identificação de *outliers*, reduziu a taxa de falsos alarmes, estabelecendo um modelo confiável de previsão da série temporal e dando credibilidade ao sistema de controle estatístico de leituras de instrumentação. (LAZZAROTTO, GRAMANI e CHAVES NETO, 2014)

APÊNDICE C (Board)

Comitê de Consultores Cíveis de Itaipu)

Nesta apresentação ao Comitê de Consultores Cíveis de Itaipu, evento que ocorre a cada 4 anos, a convite da equipe de segurança de barragem da usina, foram apresentados alguns resultados sobre a importância e a necessidade de reavaliação de valores estatístico de controle em dados de instrumentação de barragem, a forma de tratamento univariado x multivariado dos dados, a avaliação de hipóteses de gráficos de controle tradicionais: autocorrelação (independência) e a normalidade e a intenção de detectar um conjunto de instrumentos mais significativos para avaliar a variabilidade nas leituras dos instrumentos, bem como a perda da “qualidade” de previsões em modelos estatísticos quando se diminui a quantidade de dados disponível.

APÊNDICE D (SNGB)

XXX Seminário Nacional de Grandes Barragens

Neste trabalho buscou-se comparar a taxa de falsos alarmes de dados, sob controle, da leitura de instrumentos de monitoramento da barragem da usina de Itaipu, de gráficos de controle paramétricos tradicionais de Shewhart univariado e T^2 multivariado com gráficos de controle em que os limites de controle foram determinados através de simulações via técnica não paramétrica *Bootstrap*. Nas comparações, foi buscado retirar o efeito da autocorrelação na geração de falsos alarmes por meio da análise dos resíduos de modelos ARIMA e a forma de estimar a matriz de covariância em dados multivariados também foi considerada nas comparações.

O uso dos resíduos dos modelos ARIMA nos gráficos de Shewhart univariados viabiliza o uso de gráficos de controle tradicionais para os dados de instrumentação que presumem a normalidade, entretanto ainda gera uma quantidade elevada de valores fora dos limites de controle (FLC), ou seja, de falsos alarmes. Já o gráfico de controle dos resíduos de modelos ARIMA, em que os limites de controle são estabelecidos pela técnica *Bootstrap*, conseguiu obter resultados muito mais próximos da verdadeira taxa de falsos alarmes estipulada. A mesma conclusão é válida quando se compara a taxa de falsos alarmes para o gráfico T^2 de Hotelling que presume normalidade multivariada com o gráfico dos valores da estatística T^2 em que o limite superior de controle é estabelecido com base em simulações *Bootstrap*. Deve-se salientar que a tentativa de outros modelos paramétricos univariados também não produziu bons resultados.

Quando se analisa a fase II, ou seja, o conjunto total de dados disponíveis foi observado que a previsão do modelo ARIMA é boa para um horizonte médio de 3 a 4 observações seguintes, o que corresponderia, neste caso, a um período de 1 a 2 meses. Suspeita-se fortemente que se a cada nova observação for revisado o modelo de série temporal mais adequado, o resíduo gerado na previsão seguinte será menor e serão gerados menos falsos alarmes.

Mesmo assim, quando avaliamos o número de observações fora dos limites de controle (FLC) na fase II em relação à fase I, os gráficos cujos limites de controle foram calculados com base nas simulações *Bootstrap* apresentaram números menores de FLC do que em relação ao gráfico de Shewhart. Isto também reitera o fato que os gráficos tradicionais perdem sua eficiência na ausência da normalidade.

Pode ser concluído que o uso dos modelos ARIMA para retirar o efeito da autocorrelação em dados de leituras de instrumentos é fundamental, tendo em vista o elevado número de FLC que inviabiliza o controle estatístico sem seu uso. No caso multivariado, após a retirada da autocorrelação, praticamente é indiferente a forma de estimar a matriz de covariância. (LAZZAROTTO, GRAMANI e CHAVES NETO, 2015)

APÊNDICE E (CMN)

Congresso de Métodos Numéricos em Engenharia 2015

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* 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 (LAZZAROTTO *et al.*, 2015).

APÊNDICE F (IJMP)

Revista Independent Journal of Management & Production

Abstract

A high number of instruments that assess various quality characteristics of interest that have an inherent variability monitors hydroelectric plants. The readings of these instruments generate time series of data on many occasions have correlation. Each project of a dam plant has characteristics that make it unique. Faced with the need to establish statistical control limits for the instrumentation data, this article proposes an approach to multivariate statistical analysis and a model that uses principal components control charts and statistical tests T^2 and Q to explain variability and establish a method of monitoring to control future observations. An application for section E of the Itaipu hydroelectric plant is performed to validate the model. The results show that the method used is appropriate and can help identify the type of outliers, reducing false alarms and reveal instruments that have higher contribution to the variability.

(LAZZAROTTO *et al.*, 2016)

APÊNDICE G (Cadernos do IME)

Revista Cadernos do IME – Série Estatística

Resumo

Barragens de usinas hidrelétricas são avaliadas por inspeções visuais e por instrumentos de monitoramento. As leituras periódicas de um instrumento podem ser interpretadas como uma série temporal estocástica cujos valores passados fornecem informações relevantes para construção de previsões acerca de seus valores futuros, além de indicar tendências do comportamento futuro da barragem como um todo. Desta forma, é fundamental produzir previsões mais acuradas possíveis, de modo que sirvam de alerta confiável para predição de eventuais anormalidades no comportamento da barragem, permitindo a realização precoce de obras e ações de intervenção. Assim, este artigo propõe uma metodologia na qual é feita uma avaliação do desempenho por meio da previsão de séries temporais das leituras de um instrumento da barragem da usina hidrelétrica de Itaipu usando uma combinação híbrida dos previsores ARIMA-GARCH e Redes Neurais Artificiais, com a utilização da decomposição Wavelet. Os resultados mostram que a combinação híbrida proposta alcançou desempenho, em termos de acurácia, bastante superior quando comparado com o uso individual dos métodos preditivos tradicionais e de suas combinações.

(LAZZAROTTO *et al.*, 2015)

APÊNDICE H (EAAS - IJEAS)

Revista International Journal of Engineering and Applied Sciences

Resumo

This paper proposes a hybrid methodology for combining forecasts to (stochastic) time series referred to as Wavelet Linear Combination (WLC) SARIMA-ANN with Multiple Stages. Firstly, the Wavelet decomposition of level p is performed, generating (approximations of the) $p+1$ Wavelet components (WCs). Then, the WCs are individually modeled by means of a Box and Jenkins model and an artificial neural network in order to capture, respectively, plausible linear and non-linear structures of auto-dependence for, then, being linearly combined, providing hybrid forecasts for each one. Finally, all of them are linearly combined by the WLC of forecasts (to be defined). For evaluating, it was used the Box and Jenkins (BJ) models, artificial neural networks (ANN), and its traditional Linear Combination (LC1) of forecasts; and ANN integrated with the Wavelet decomposition (ANN-WAVELET), BJ model integrated with the Wavelet decomposition (BJ-WAVELET), and its conventional Linear Combination (LC2) of forecasts. All predictive methods applied to the monthly time series of average flow of the Itaipu plant, located in Foz do Iguaçu, Brazil. In all analysis, the proposed hybrid methodology has provided higher predictive performance than the other ones.

(LAZZAROTTO *et al.*, 2015)