

**UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE
AUTOMAÇÃO E SISTEMAS**

André Ambrósio Boechat

**ALGORITMOS PARA A DETECÇÃO DE *DRIFTING* EM
SENSORES DE FUNDO DE POÇO**

Florianópolis

2012

André Ambrósio Boechat

**ALGORITMOS PARA A DETECÇÃO DE *DRIFTING* EM
SENSORES DE FUNDO DE POÇO**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas para a obtenção do Grau de Mestre em Engenharia de Automação e Sistemas.

Orientador: Prof. Dr. Ubirajara Franco Moreno

Florianópolis

2012

André Ambrósio Boechat

**ALGORITMOS PARA A DETECÇÃO DE *DRIFTING* EM
SENSORES DE FUNDO DE POÇO**

Esta Dissertação foi julgada aprovada para a obtenção do Título de “Mestre em Engenharia de Automação e Sistemas”, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Automação e Sistemas.

Florianópolis, 29 de Agosto 2012.

Prof. Dr. Jomi Fred Hübner
Coordenador do Curso

Banca Examinadora:

Prof. Dr. Ubirajara Franco Moreno
Orientador

Prof. Dr. Marcelo Ricardo Stemmer

Prof. Dr. Max Hering de Queiroz

Prof. Dr. Rodrigo Ivan Goytia Mejia

AGRADECIMENTOS

Ao meu orientador prof. Ubirajara F. Moreno, pela troca de ideias, direcionamento do trabalho e por toda cordialidade.

Ao grupo de trabalho do projeto de pesquisa com o Cenpes/Petrobras, pelo importante compartilhamento de conhecimentos.

Ao Cenpes/Petrobras, pelo suporte financeiro e fornecimento de informações cruciais para o desenvolvimento do trabalho.

Aos grandes amigos que fiz durante esta jornada.

In God we trust, all others bring data.

William Edwards Deming

RESUMO

Apresentam-se neste trabalho o desenvolvimento e a implementação de sistemas de monitoramento e validação de sensores de poços de petróleo, utilizando técnicas de aprendizado de máquina para a construção de modelos empíricos baseados em histórico de dados. Tais sistemas são fundamentados na redundância analítica dos sensores e o foco principal é a detecção de possíveis desvios nas leituras realizadas por esses dispositivos. Basicamente, o processo de validação é composto por três etapas: a predição das atuais leituras de sensores a partir de seus históricos de dados, a comparação entre as predições geradas e as correspondentes leituras dos sensores e, finalmente, a detecção de mudanças estatísticas nas propriedades dos resultados dessa comparação. Para verificar a aplicabilidade dos sistemas desenvolvidos, são realizados ensaios com diferentes conjuntos de dados: um conjunto de dados gerado a partir da simulação do modelo de um poço de petróleo e um conjunto de dados coletados de sensores de um poço real. De acordo com os resultados obtidos e as métricas de desempenho adotadas, o emprego de modelos empíricos baseados em histórico para a validação de sensores de poços de petróleo é promissor.

Palavras-chave: validação de sensores, desvios, modelos regressivos, monitoramento de desempenho, manutenção baseada em condição, monitoramento de poços de petróleo.

ABSTRACT

This work demonstrates the development and implementation of monitoring and validation systems for oil well sensors, using machine learning techniques to build data-driven models of the behaviour of correlated sensors. These systems are based on the analytical redundancy of the sensors and the main objective is the calibration drift detection and isolation. Basically, the validation process is composed of three steps: the prediction of the sensors values given by the data-driven models, the comparison between the prediction and the actual sensor value, generating residuals, and, finally, the detection of abnormal changes in the statistical properties of the residuals. To verify the applicability of the developed systems, tests with two different datasets are performed: one dataset generated by the simulation of an oil well model and one dataset collected from sensors of a real oil well. Considering the obtained results and the performance metrics adopted, the use of data-driven models for validation of oil well sensors is promising.

Keywords: sensor validation, drift, regressive models, performance monitoring, condition based maintenance, oil well monitoring.

LISTA DE FIGURAS

Figura 1	Esquema de funcionamento de sistemas baseados em modelos empíricos para monitoramento do desempenho de sensores. Adaptado de Ma e Jiang (MA; JIANG, 2011).	28
Figura 2	Arquitetura conceitual de sistemas baseados na arquitetura OSA-CBM (Penn State University; The Boeing Company, 2006).	33
Figura 3	Arquitetura conceitual de sistemas baseados na arquitetura OSA-CBM. Adaptada de Souza (SOUZA, 2008).	34
Figura 4	Relações entre as arquiteturas OSA-CBM e OSA-EAI. Adaptada de Swearingen (SWEARINGEN et al., 2007).	35
Figura 5	Arquitetura do sistema de validação de sensores baseada no modelo OSA-CBM.	36
Figura 6	Zona de insensibilidade e penalização no modelo SVM (SMOLA; SCHÖLKOPF, 2004).	47
Figura 7	Diagrama do processo de monitoramento pelo Sistema 1.	55
Figura 8	Diagrama do processo de monitoramento de um único sensor pelo Sistema 2.	56
Figura 9	Diagrama do processo de monitoramento pelo Sistema 3.	57
Figura 10	Esquema do poço usado para gerar os dados de simulação.	60
Figura 11	Dados de variáveis gerados pela simulação de um poço de petróleo.	61
Figura 12	Relação entre o MSE médio e diferentes valores de largura de banda h e vetores de memória n_m para o modelo AAKR na predição do conjunto de dados de otimização gerados por simulação.	62
Figura 13	Predições dos sistemas para o conjunto de dados de testes gerados por simulação, com inserção de desvios na variável PT_g . Linhas vermelhas são as predições, linhas verdes são as leituras das variáveis (sujeitas a falhas) usadas como entradas dos sistemas e linhas azuis são os valores verdadeiros das variáveis (diferentes das linhas verdes apenas na ocorrência de falhas). Todos os dados foram escalados no intervalo $[0, 1]$	64
Figura 14	Resíduos gerados pelo Sistema 1 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados de simulação.	65
Figura 15	Resíduos gerados pelo Sistema 2 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados	

de simulação.	65
Figura 16 Resíduos gerados pelo Sistema 3 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados de simulação.	66
Figura 17 Dados coletados por sensores de um poço de petróleo real, amostrados à taxa de uma amostra por minuto.	67
Figura 18 Predições dos sistemas para o conjunto de dados reais de testes, com inserção de desvios nas leituras do PDG. Linhas vermelhas são as predições, linhas verdes são as leituras dos sensores (sujeitas a falhas) usadas como entradas dos sistemas e linhas azuis são as leituras originais dos sensores (diferentes das linhas verdes apenas na ocorrência de falhas). Todos os dados foram escalados no intervalo $[0, 1]$	69
Figura 19 Resíduos gerados pelo Sistema 1 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.	70
Figura 20 Resíduos gerados pelo Sistema 2 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.	71
Figura 21 Resíduos gerados pelo Sistema 3 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.	71

LISTA DE TABELAS

Tabela 1	Configuração dos modelos SVM construídos com conjunto de dados de treinamento gerados por simulação.	61
Tabela 2	Valores médios das métricas de desempenho para os sistemas de monitoramento e validação de sensores: (E_a), autossensibilidade (S_A), sensibilidade cruzada (S_C) e erro de predição (E_p).	62
Tabela 3	Configurações do algoritmo SPRT para o conjunto de dados de teste gerados por simulação.	63
Tabela 4	Configuração dos modelos SVM construídos com conjunto de dados reais de treinamento.	68
Tabela 5	Valores médios das métricas de desempenho para os sistemas de monitoramento e validação de sensores: (E_a), autossensibilidade (S_A), sensibilidade cruzada (S_C) e erro de predição (E_p).	68
Tabela 6	Configurações do algoritmo SPRT para o conjunto de dados reais de teste.	70

LISTA DE ABREVIATURAS E SIGLAS

CBM	Manutenção baseada na condição	25
OLM	On-line Monitoring	25
AANN	Redes neurais artificiais auto-associativas	28
SPRT	Sequential Probability Ratio Test	28
MSET	Multivariate State Estimation Techniques	28
SVM	Support Vector Machines	28
AAKR	Regressão kernel auto-associativa	28
KF	Filtro de Kalman	29
OSA-CBM	Open System Architecture Condition-Based Maintenance ...	32
OSA-EAI	<i>Open System Architecture for Enterprise Application Integration</i>	35
RBF	Função de base radial	48
MSE	Erro quadrático médio	50

LISTA DE SÍMBOLOS

p	Número de sensores cujas leituras compõem uma amostra de dados	27
\mathbf{r}	$[r_1, \dots, r_p]$, vetor que representa uma amostra de dados de p sensores	27
$\hat{\mathbf{r}}$	$[\hat{r}_1, \dots, \hat{r}_p]$, estimativa de \mathbf{r} gerada por um modelo regressivo	27
\mathbf{d}	$[d_1, \dots, d_p]$, vetor de resíduos entre \mathbf{r} e $\hat{\mathbf{r}}$	27
n_m	Quantidade de amostras de dados usados para construção dos modelos regressivos	27
\mathbf{X}	Matriz de dimensão $n_m \times p$ de histórico de amostras de treinamento do modelo regressivo AAKR.	27
y	$f(\mathbf{r}) = y$, saída ou predição de um modelo inferencial a partir de um vetor \mathbf{r}	43
u_i	Distância Euclidiana entre o vetor de memória $\mathbf{x}^{(i)}$ e um vetor de entrada \mathbf{r}	44
K	Função kernel	44
h	Largura de banda da função kernel Gaussiana	44
\bar{w}_i	Peso relativo ao vetor u_i e à largura de banda h	44
$v^{(k)}$	$d^{(k)} = d^{(k-1)} + v^{(k)}$, ruído Gaussiano na dinâmica dos desvios do filtro de Kalman	51
σ_v	$v^{(k)} \sim \mathcal{N}(0, \sigma_v)$, variância do ruído Gaussiano do modelo da dinâmica dos desvios incorporado no filtro de Kalman	51
$q^{(k)}$	$z = d^{(k)} + q^{(k)}$, ruído Gaussiano na equação de observação do filtro de Kalman	51
σ_q	$q^{(k)} \sim \mathcal{N}(0, \sigma_q)$, variância do ruído Gaussiano na equação de observação do filtro de Kalman	51
E_a	Erro de acurácia	57
S_A	Autosensibilidade	57
S_C	Sensibilidade cruzada	58
E_p	Erro de predição no caso de ocorrência de desvios nas entradas do sistema	58
σ_N	Desvio padrão do ruído branco adicionado aos dados de uma variável de simulação.	60
σ_S	Desvio padrão dos dados de uma variável de simulação.	60

SUMÁRIO

1 INTRODUÇÃO	25
1.1 DEFINIÇÃO DO PROBLEMA E ABORDAGENS EXISTENTES	27
1.2 OBJETIVOS E CONTRIBUIÇÕES	29
1.3 ESTRUTURA DA DISSERTAÇÃO	30
2 ESTRUTURA DE SISTEMAS DE MONITORAMENTO E VALIDAÇÃO DE SENSORES	31
2.1 OSA-CBM	32
2.2 VALIDAÇÃO DE SENSORES E O MODELO OSA-CBM	35
3 PROCESSAMENTO DE SINAIS PARA DETECÇÃO E CORREÇÃO DE DRIFT	39
3.1 PRÉ-PROCESSAMENTO DOS SINAIS	39
3.1.1 Limpeza dos Dados	39
3.1.2 Normalização	40
3.1.3 Seleção das Variáveis de Entrada	40
3.1.4 Seleção das Amostras de Treinamento	41
3.2 PROCESSAMENTO DOS SINAIS	42
3.2.1 Seleção do Tipo de Estrutura	42
3.2.2 Técnicas de Regressão	43
3.2.2.1 AAKR	43
3.2.2.1.1 Seleção dos vetores de memória	45
3.2.2.2 SVM	46
3.2.3 Construção e Otimização dos Modelos	49
3.3 PÓS-PROCESSAMENTO DOS SINAIS	50
3.3.1 Filtro de Kalman	51
3.4 MONITORAMENTO DE CONDIÇÃO	52
3.4.1 SPRT	53
4 APLICAÇÃO DOS SISTEMAS IMPLEMENTADOS A DADOS DE POÇOS DE PETRÓLEO	55
4.1 SISTEMAS DE VALIDAÇÃO DE SENSORES	55
4.1.1 Sistema 1 — AAKR-SPRT	55
4.1.2 Sistema 2 — SVM-SPRT	56
4.1.3 Sistema 3 — AAKR-KF-SPRT	56
4.2 MÉTRICAS DE DESEMPENHO	57
4.3 ENSAIOS	58
4.3.1 Ensaios com Dados de Simulação	59
4.3.2 Ensaios com Dados Reais	67
5 CONCLUSÕES	73

Referências Bibliográficas 75

1 INTRODUÇÃO

Na indústria petrolífera, o emprego de sensores nos poços de produção possui papel fundamental tanto na segurança quanto no desempenho das operações. Os dados coletados fornecem medidas das condições de funcionamento de equipamentos e de parâmetros diretamente relacionados com as ações de controle, otimização da produção e monitoramento. O operador da planta interpreta as medições baseado na sua experiência e conhecimento do sistema, e toma decisões que podem ser cruciais durante situações de emergência. Entretanto, devido às condições degradantes a que ficam submetidos, os sensores de poços normalmente apresentam algum tipo de falha ou *drift* (desvio nas medidas) durante o período produtivo de um poço. Trocas ou reparos desses sensores raramente ocorrem, mesmo quando se sabe que as medições estão incorretas, uma vez que possuem difícil acesso e alto custo de manutenção (AGGREY; DAVIES, 2007), principalmente em plataformas *offshore*. Por isso, a validação das leituras dos sensores e a determinação de seus estados de funcionamento são competências altamente desejáveis.

Em diversos setores da indústria, a validação de sensores é realizada através dos métodos tradicionais de manutenção, os quais envolvem a calibração periódica dos instrumentos. Várias técnicas de calibração periódica necessitam de paradas dos processos de produção e/ou de se retirar de operação os instrumentos de medição. Entretanto, como apresentado por Hines et al. (HINES et al., 2008a), estudos recentes mostraram que menos de 5% da calibração manual realizada é sequer necessária. Além disso, a confiabilidade de um instrumento pode ser afetado de forma adversa pelas intervenções manuais.

Por essas e outras razões, como a competitividade de mercado, tem crescido a procura por estratégias menos invasivas e mais eficientes. Técnicas de manutenção baseadas na condição (CBM — *Condition Based Maintenance*) tendem à manutenção ótima, pois a condição de funcionamento dos instrumentos é monitorada durante a operação da planta e recalibrações físicas são realizadas apenas quando uma degradação do desempenho é detectada. Na literatura, esses métodos têm sido chamados de monitoramento *on-line* de calibração (OLM — *On-line Monitoring*).

O monitoramento de calibração baseia-se essencialmente em estimar as medidas corretas que deveriam ser realizadas pelos sensores, comparando-as com as reais medidas efetuadas pelos mesmos. Para isso, existem duas abordagens atualmente utilizadas: redundância por *hardware* e redundância analítica (MA; JIANG, 2011).

Na redundância por *hardware*, sensores redundantes são utilizados

para medirem uma mesma variável, possibilitando formas mais simples e intuitivas de estimação dos valores corretos das medições (pela média das medições, por exemplo). Entretanto, além da necessidade de aquisição de sensores extras, a redundância por *hardware* apresenta dificuldades na detecção de sensores descalibrados que apresentam *drift* na mesma direção. Além disso, segundo Eck et al. (ECK et al., 1999), no caso dos sensores de fundo de poço, componentes redundantes ocupam um espaço limitado, e consomem uma energia preciosa.

A outra abordagem, por redundância analítica, consiste na estimação das medidas dos sensores baseando-se em outras medidas correlacionadas disponíveis no sistema. Existem duas formas principais de modelagem dessas correlações: a modelagem por equações físicas que descrevem as interações entre as variáveis e modelagem baseada em dados. Os modelos baseados em equações físicas, apesar de serem normalmente muito precisos, necessitam de significativos esforços de engenharia e são muito sensíveis à mudanças ou degradações não previstas no sistema. Os modelos empíricos baseados em dados, ou histórico, também se baseiam em medidas correlacionadas dentro do sistema, mas essas correlações ficam implícitas, capturadas por técnicas de inteligência artificial e aprendizado de máquinas durante a análise de dados de medições livres de falhas, coletados durante situações normais de operação da planta. Apesar de serem normalmente limitados a trabalharem em pontos de operação da planta semelhantes aos quais foram treinados, os modelos empíricos possuem diversos benefícios práticos, como: aplicabilidade livre de contexto, ou seja, a essência dos modelos pode ser aplicada a quaisquer tipos de sistemas, sem a necessidade de conhecimentos profundos sobre este; simplicidade de desenvolvimento, uma vez que não existe a necessidade de modelos explícitos, o que torna-se um atrativo quando se trata de sistemas complexos; e flexibilidade de configuração, facilitando a configuração dos modelos para atender a novos requisitos de desempenho ou a mudanças na própria planta.

Esta dissertação tem como foco o desenvolvimento e implementação de um sistema de monitoramento do desempenho de sensores de poços de petróleo, cujas técnicas são baseadas em técnicas de aprendizado de máquina e modelos empíricos construídos com histórico de dados. Como as técnicas baseadas em modelos empíricos são livres de contexto, este trabalho possui potencial aplicação para diferentes setores industriais.

1.1 DEFINIÇÃO DO PROBLEMA E ABORDAGENS EXISTENTES

Como apresentado por Mauro Oliveira (Mauro Vitor de Oliveira, 2005), a validação de sinal pode ser definida como a detecção, isolamento e caracterização de sinais falhos. Em sistemas OLM, a validação de sinais também é referida como a identificação de falhas em sensores acompanhada da estimativa ou predição das medições corretas, durante o funcionamento da planta ou processo.

Os sensores de poços são passíveis a vários tipos de falhas, como mudanças abruptas, polarização (*bias*), picos, etc. O principal tipo de falha tratado neste trabalho é o *drift* ou desvio de medição. Na literatura, o *drift* é normalmente definido como um desvio lento, contínuo ou incremental, das medições de um sensor ao longo do tempo.

A abordagem por modelos empíricos baseados em dados normalmente assume uma premissa fundamental: as variáveis medidas pelos sensores são correlacionadas, enquanto as possíveis falhas nesses instrumentos são descorrelacionadas. Essa premissa implica que existem diferenças entre as correlações dos dados gerados em situações normais e as dos dados gerados em situações de falhas. Os modelos são normalmente desenvolvidos utilizando histórico de medições livres de erros, capturando a correlação verdadeira entre as variáveis. Depois de construídos, esses modelos conseguem gerar estimativas das medições corretas dos sensores, as quais são comparadas com as medições reais com o intuito de se identificar e isolar possíveis falhas. Casos de sucesso no emprego de tais técnicas na indústria já foram reportados, como é o caso das plantas de energia nuclear (MA; JIANG, 2011).

Normalmente, os sistemas de monitoramento do desempenho de sensores baseados em modelos empíricos segue o esquema apresentado na Fig. 1. Supondo uma matriz de dados de treinamento $\mathbf{X} \in \mathbb{R}^{n_m \times p}$, onde n_m é o número de amostras de treino e p é a quantidade de variáveis amostradas, um modelo empírico f é treinado usando \mathbf{X} . Quando novas medições $\mathbf{r} \in \mathbb{R}^{1 \times p}$ tornam-se disponíveis, estimações de \mathbf{r} são obtidas por $\hat{\mathbf{r}} = f(\mathbf{r})$ e resíduos são gerados como $\mathbf{d} = \mathbf{r} - \hat{\mathbf{r}}$. A ocorrência de *drift* nas medições causa mudanças nas relações entre as variáveis de \mathbf{r} , o que resulta em mudanças estatísticas anormais nos resíduos. Então, avaliando estatisticamente os resíduos podem-se estabelecer as condições de operação ou a saúde dos sensores (MA; JIANG, 2011).

No trabalho de Hines e Uhrig (HINES; UHRIG, 1998), um sistema de monitoramento do desempenho de sensores foi implementado utilizando basicamente uma rede neural auto-associativa (AANN) como modelo empírico e o algoritmo SPRT (*Sequential Probability Ratio Test*) para a análise das propriedades estatísticas dos resíduos. O sistema foi avaliado em dados

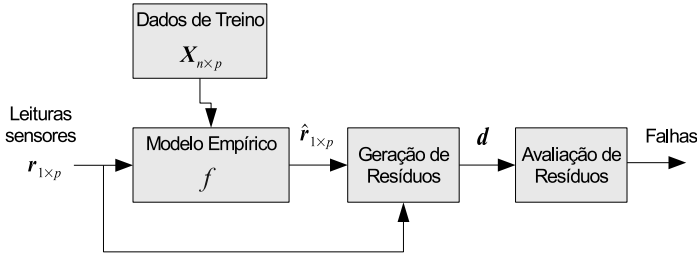


Figura 1 – Esquema de funcionamento de sistemas baseados em modelos empíricos para monitoramento do desempenho de sensores. Adaptado de Ma e Jiang (MA; JIANG, 2011).

de plantas nucleares, envolvendo problemas como *drifts* e erros grosseiros nos sensores. Apesar dos bons resultados, o treinamento do modelo inicial e o retreinamento para mudanças significativas nas condições de operação da planta possuem alto custo computacional.

Gribok, Hines e Uhrig (GRIBOK; HINES; UHRIG, 2000) realizaram uma comparação entre diferentes técnicas estatísticas para a predição de dados de plantas nucleares. Entre as técnicas utilizadas, encontram-se redes neurais, mínimos quadrados ordinários, regressão kernel, MSET (*Multivariate State Estimation Techniques*) e SVM (*Support Vector Machines*). Os resultados apontaram as técnicas baseadas em kernel como promissoras, principalmente a SVM pela sua inerente regularização e estabilidade das predições.

Zavaljevski e Gross (ZAVAJJEVSKI; GROSS, 2000) realizaram a validação de sensores de reatores nucleares utilizando uma combinação de kernels MSET e do método SVM. As predições eram comparadas com os dados dos sensores para a formação de resíduos, cujas propriedades estatísticas eram analisadas usando SPRT para a detecção de falhas.

Garvey et al. (GARVEY et al., 2007) apresentaram os resultados da aplicação da técnica de regressão kernel auto-associativa (AAKR) a conjuntos de sensores de plantas nucleares. Diferentes métricas foram apresentadas para a avaliação de desempenho dos modelos AAKR e a detecção de desvios foi realizada a partir dos intervalos de confiança construídos pela estimação da incerteza dos modelos. Um dos ganhos de se estimar as incertezas dos modelos é a capacidade de se indicar, a partir de uma métrica, o menor desvio detectável nos sensores.

Aggrey e Davies (AGGREY; DAVIES, 2007) apresentaram um estudo sobre o uso de redes neurais para a predição de dados de sensores de fundo de poços de petróleo e estimação de parâmetros de reservatórios. Vários ganhos que se poderiam obter para a indústria de petróleo, advindos do emprego de modelos baseados em histórico (*soft sensing*), foram citados. Os experimen-

tos consistiram no emprego de uma rede neural inferencial para se estimar a pressão de fundo de poço considerando medidas de outros sensores correlacionados, cujas leituras correspondiam à temperatura e pressão de diferentes pontos do poço. Os resultados do modelo empírico foram consideradas muito precisas ($\approx 0.1\%$ de erro) e robustas a falhas. Entretanto, os testes incluíram apenas falhas no sensor estimado pela rede neural, cujas medidas não são usadas pela rede durante as predições. Nenhuma forma de detecção automática de falhas de sensores foi utilizada.

Takruri, Rajasegarar e Challa (TAKRURI; RAJASEGARAR; CHALLA, 2008) propuseram um novo algoritmo para correção de medições de redes de sensores sem fio. O algoritmo foi concebido para trabalhar de forma descentralizada, sendo executado por cada sensor correlacionado de uma rede. Cada sensor recebe as leituras de sensores correlacionados, com as quais infere a própria leitura utilizando um modelo SVM. A leitura inferida e a leitura real são utilizadas por um filtro de Kalman (KF) que possui um modelo genérico de *drift*, gerando uma estimação do valor verdadeiro da variável mensurada pelo sensor. Em trabalhos mais recentes de Takruri, Challa e Yunis (TAKRURI; CHALLA; YUNIS, 2009), o filtro de Kalman foi substituído por um *Unscented Kalman Filter*, resultando na redução do *bias* normalmente gerado pela formulação anterior.

Neste trabalho, as técnicas de predição escolhidas são AAKR e SVM. Comparado aos demais modelos, os modelos AAKR possuem implementação computacional simples; os modelos SVM possuem muitos relatos de sucesso na literatura e estrutura inferencial, gerando estimativas das leituras de um sensor monitorado sem a necessidade das respectivas leituras reais. A análise estatística dos resíduos é realizada pelo algoritmo SPRT, largamente utilizado na literatura para testes sequenciais de hipóteses probabilísticas. Utilizando tais técnicas, a implementação do sistema de monitoramento de desempenho e detecção de desvios em sensores é avaliada pela consistência e qualidade das predições e pela capacidade de detecção e isolamento de desvios.

1.2 OBJETIVOS E CONTRIBUIÇÕES

O principal objetivo desta dissertação é apresentar o esquema de funcionamento de um sistema para detecção e correção de desvios em sensores dentro de uma arquitetura mais genérica para sistemas de monitoramento e manutenção baseado na condição. Para se atingir esse objetivo principal, os seguintes objetivos são incluídos como etapas intermediárias:

- Desenvolver um sistema de detecção/correção de desvios em sensores baseado em modelos empíricos e técnicas de aprendizado de máquina;

- Ensaiar diferentes técnicas de modelagem empírica e configurações do sistema;
- Ensaiar e avaliar o sistema desenvolvido com dados de poços de produção de petróleo.

Como principais contribuições deste trabalho, podem-se destacar:

- Aplicação ao cenário de poços de produção de petróleo técnicas recentes de modelagem empírica baseada em histórico de dados;
- Apresentação de uma estrutura modular para implementação do sistema de validação de sensores dentro de uma arquitetura mais geral para monitoramento e manutenção baseado na condição;
- Emprego de um modelo de estimação de *drift* ao resíduo gerado entre um modelo empírico de predição auto-associativo e as leituras dos sensores.

1.3 ESTRUTURA DA DISSERTAÇÃO

No Capítulo 2, é apresentada a estrutura do sistema de validação de sensores e estabelecida uma relação entre essa estrutura e uma arquitetura mais geral para sistemas CBM. Em seguida, no Capítulo 3, são apresentadas as técnicas de modelagem empírica e de monitoramento, ressaltando suas abordagens ao problema de validação de sensores. A descrição do sistema de validação de sensores implementado e sua avaliação em diferentes cenários de ensaios são apresentadas no Capítulo 4. Finalizando o trabalho, o Capítulo 5 apresenta as conclusões sobre o desempenho do sistema de validação e dos modelos empíricos desenvolvidos e as perspectivas de trabalhos futuros.

2 ESTRUTURA DE SISTEMAS DE MONITORAMENTO E VALIDAÇÃO DE SENSORES

Em esquemas de manutenção baseados em CBM, as condições de funcionamento de equipamentos e sistemas são monitoradas com o intuito de otimizar as ações de manutenção. As manutenções preventivas deixam de ser periódicas, passando a serem agendadas de forma dinâmica, de acordo com evidências de real necessidade. A implementação de estratégias desse tipo pode reduzir o tempo de inatividade de plantas de produção, melhorar o desempenho de sistemas e o gerenciamento de recursos humanos limitados, prolongar a vida útil de equipamentos, reduzir custos e tornar a produção gradualmente mais efetiva.

A tendência atual de utilização de centros integrados remotos, localizados em terra, para suporte à operação, manutenção e otimização de unidades marítimas de produção de petróleo tem incentivado a utilização de ferramentas de manutenção baseadas na condição. Esses centros possuem equipes altamente capacitadas que, munidas com as informações certas e no momento adequado, são capazes de avaliar o desempenho e realizar diagnósticos sobre equipamentos críticos (compressores, bombas, turbinas, etc). Ferramentas CBM podem garantir o enriquecimento dos dados provenientes de plataformas ao mesmo tempo que reduz a quantidade de informações a serem analisadas pelas equipes, causando um consequente aumento da segurança e confiabilidade nas tomadas de decisão, além de reduzir custos materiais e humanos.

Apesar dos benefícios obtidos com a estratégia CBM, sérias dificuldades são frequentemente encontradas para implementá-la de forma plena (CAMPOS, 2009). A quantidade de dados coletados torna-se normalmente muito grande. Pode haver a necessidade de coleta de dados provindos de sistemas geograficamente dispersos. Os dados precisam ser integrados para proverem informações úteis. Com o passar do tempo, pode ser necessário a aquisição de dados de novas fontes e integrá-los com o restante para se obter mais informações significativas. Finalmente, torna-se indispensável a disponibilidade de conhecimentos especialistas para converter os dados gerados em informações úteis para manutenção.

No caso da indústria de petróleo, a implementação de um sistema CBM que envolvesse todas as etapas do processo de produção se tornaria altamente complexo, mesmo utilizando-se técnicas simples para o processamento de sinais, monitoramento e detecção. Daí vem a necessidade de uma metodologia de desenvolvimento que permitisse a implementação do sistema de manutenção de forma bem estruturada, dentro de uma arquitetura flexível e

robusta. A padronização da especificação de uma interface dentro da comunidade CBM poderia, idealmente, direcionar os fornecedores de soluções CBM a produzirem componentes de hardware e software intercambiáveis. Múltiplos desenvolvedores poderiam atuar na solução de um mesmo sistema. Entre os potenciais benefícios de um padrão não proprietário, robusto e largamente adotado, podem-se citar:

- facilidade para atualização de componentes de sistema;
- aumento do número de fornecedores, resultando em mais opções de tecnologia;
- desenvolvimento tecnológico mais rápido;
- redução de custos e preços.

A seguir são apresentadas uma breve descrição sobre a arquitetura de sistemas de manutenção baseados em CBM adotada como diretriz, a OSA-CBM, e uma proposta de configuração de um sistema de validação de sensores dentro dessa arquitetura.

2.1 OSA-CBM

Recentemente, uma associação entre indústrias, fabricantes, universidades e a Marinha Norte-Americana desenvolveu uma arquitetura aberta para sistemas de manutenção baseados em CBM, conhecida como OSA-CBM (*Open System Architecture Condition-Based Maintenance*). O foco principal dessa aliança foi desenvolver uma arquitetura para sistemas CBM que facilitasse a interoperabilidade entre componentes de *hardware* e *software*.

No OSA-CBM, a arquitetura de um sistema CBM é dividida em diferentes camadas funcionais, cujos conteúdos devem ser implementados seguindo as interfaces definidas pelo padrão. Atualmente, as camadas funcionais, ou módulos, somam um total de seis, seguindo a norma ISO-13374. As camadas são: aquisição de dados, processamento de sinais, monitoramento de condição, avaliação de saúde (diagnóstico), a de prognóstico e a de suporte a tomada de decisão. Cada camada possui a capacidade de requisitar dados de outras camadas funcionais quando necessário, apesar do fluxo de dados normalmente ocorrer entre as camadas adjacentes. Na Fig. 2 é apresentado um esquema da organização dos módulos da arquitetura.

Os primeiros três blocos são específicos para determinadas tecnologias (monitoramento de vibração, temperatura etc.) e provêm as seguintes funções (Penn State University; The Boeing Company, 2006):

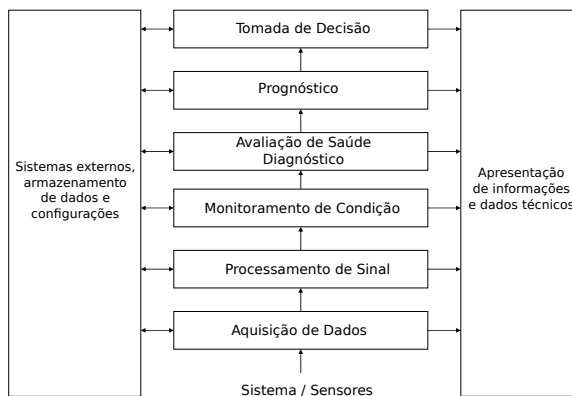


Figura 2 – Arquitetura conceitual de sistemas baseados na arquitetura OSA-CBM (Penn State University; The Boeing Company, 2006).

Aquisição de dados: conversão da saída de um transdutor ou sensor em um dado digital que represente uma medida física ou uma informação relacionada (como tempo, qualidade dos dados etc.).

Processamento de sinal: análises de sinal, cômputo de indicadores e derivação de leituras virtuais de sensores.

Monitoramento de condição: criação e manutenção de perfis de comportamento normal do sistema, procura por anomalias a cada novo dado adquirido e determinação da zona de anormalidade no qual o sistema pode se encontrar. Em resumo, o bloco gera indicadores, ou sintomas, do estado de funcionamento.

Os três blocos restantes combinam conhecimentos especialistas com tecnologias de monitoramento para se determinar o estado de funcionamento corrente do sistema, realizar previsões de falhas e recomendações de possíveis ações a serem tomadas por operadores e mantenedores do sistema. Podem-se citar as funções:

Avaliação de saúde (diagnóstico): diagnóstico das falhas e situações anômalas e avaliação da condição de operação do sistema de acordo com todas as informações de estado.

Prognóstico: previsão dos estados de operação, modos de falhas e vida útil de equipamentos/sistemas baseado na condição de operação corrente e no planejamento de carga do sistema.

Tomada de decisão: fornecer possíveis ações relativas à manutenção e a mudanças de operação necessárias para otimizar a vida útil e desempenho do processo, equipamento ou sistema.

A Figura 3, adaptada de Souza (SOUZA, 2008), ilustra um exemplo mais prático da arquitetura de um sistema de manutenção CBM, apresentando algumas possíveis técnicas presentes em cada uma das camadas. A arquitetura é dividida em duas aplicações, aplicação servidor e aplicação cliente, conectadas entre si por uma rede, onde compartilham informações. A aplicação servidor é responsável por prover as funcionalidades das seis camadas do modelo OSA-CBM, realizando a captura dos dados e o processamento inteligente. A aplicação cliente é responsável pela apresentação, através de uma interface gráfica, de dados técnicos e resultados provenientes do processamento realizado pela aplicação servidor. Possíveis aplicações da arquitetura OSA-CBM na indústria de petróleo são apresentadas por Boechat, Moreno e Teixeira (BOECHAT; MORENO; TEIXEIRA, 2011).

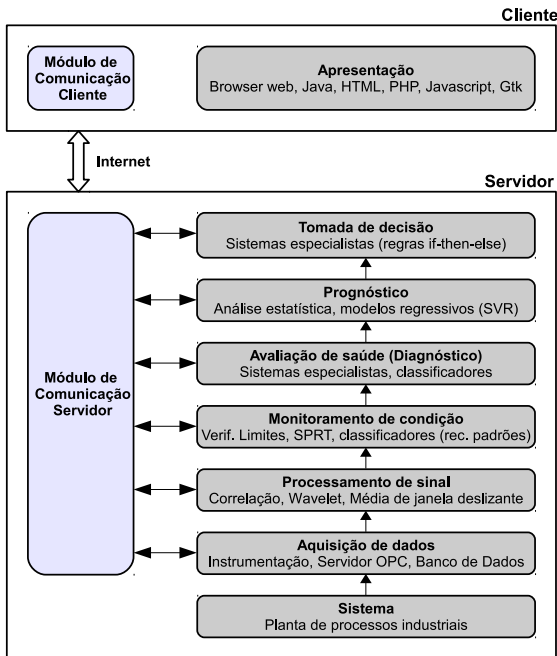


Figura 3 – Arquitetura conceitual de sistemas baseados na arquitetura OSA-CBM. Adaptada de Souza (SOUZA, 2008).

A OSA-CBM também está relacionada a outra arquitetura, a OSA-

EAI (*Open System Architecture for Enterprise Application Integration*). A OSA-EAI define estruturas de dados para armazenamento e movimentação de informações coletivas sobre todos os aspectos dos equipamentos dentro das atividades empresariais. Isso inclui a configuração física de plataformas assim como a confiabilidade, condição e manutenção de plataformas, sistemas e subsistemas. Vários elementos definidos pela OSA-EAI são utilizados pela OSA-CBM. Segundo Swearingen (SWEARINGEN et al., 2007), espera-se que no futuro a OSA-CBM esteja totalmente mapeada dentro da OSA-EAI, sem que seja possível distingui-las. A Figura 4 ilustra a atual relação entre a OSA-CBM e a OSA-EAI.

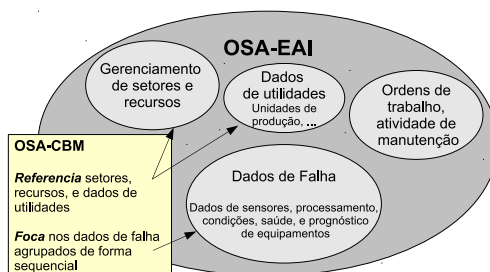


Figura 4 – Relações entre as arquiteturas OSA-CBM e OSA-EAI. Adaptada de Swearingen (SWEARINGEN et al., 2007).

2.2 VALIDAÇÃO DE SENSORES E O MODELO OSA-CBM

Considerando a arquitetura OSA-CBM, um sistema básico de validação de sensores poderia ser composto por quatro diferentes módulos e uma base de dados de histórico, como ilustrado na Fig. 5. As principais funcionalidades presentes em cada módulo são:

Aquisição de dados: composta basicamente por sensores e transdutores, responsáveis pela medição de grandezas físicas e pela transformação dessas medidas para dados digitais. Contando com um sistema de banco de dados, as medidas são armazenadas para futuras análises de histórico.

Processamento: responsável pela extração de informações relevantes sobre a condição de funcionamento dos sensores a partir das medições realizadas pelos sensores. Este módulo pode ser subdividido em pelo menos três etapas:

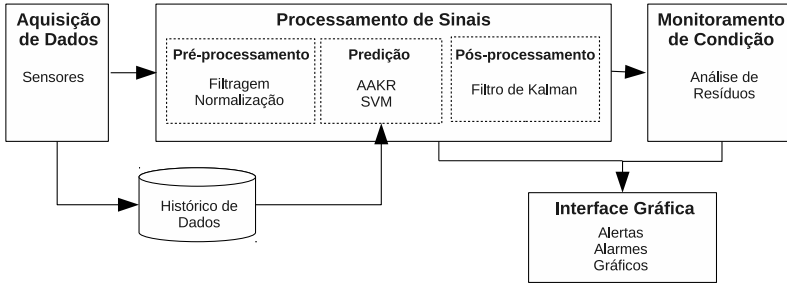


Figura 5 – Arquitetura do sistema de validação de sensores baseada no modelo OSA-CBM.

Pré-processamento: consiste na preparação e condicionamento dos dados aos algoritmos de regressão.

Predição: responsável pela construção de modelos regressivos a partir do histórico de medidas dos sensores, predição de sinais e geração de resíduos.

Pós-processamento: não intrinsecamente necessária, realiza a estimação de desvios nos sinais dos sensores para maior robustez dos modelos regressivos.

Monitoramento de condição: realiza a tarefa de monitorar os indicadores de falhas nos sensores a partir da detecção de mudanças nas propriedades estatísticas dos resíduos ou das estimativas de desvio. Alertas e alarmes podem ser gerados para indicarem falhas em sensores específicos.

Interface gráfica: módulo de apresentação de resultados e indicadores para os usuários, realiza a tarefa de apresentar as informações relevantes para cada grupo de usuários, como o de manutenção, operação da planta ou o grupo responsável por tomadas de decisões estratégicas de negócio.

No caso da arquitetura apresentada na Fig. 5, o diagnóstico de ocorrência de desvios deve ser realizado implicitamente pelo usuário a partir das indicações de mudanças nas propriedades estatísticas dos resíduos apresentadas pela camada de monitoramento da condição. Em casos mais gerais, onde a detecção de desvios faria parte de um sistema de manutenção mais completo, a implementação de uma rotina para o diagnóstico (como um conjunto de regras do tipo *if-then-else* sobre as propriedades estatísticas dos resíduos),

dentro do módulo de avaliação de saúde, seria de maior importância e mais coerente com a proposta da arquitetura.

Vale destacar ainda uma das principais vantagens da adoção de uma arquitetura padronizada como a OSA-CBM: a facilidade de inserção de novos algoritmos para cada um dos módulos do sistema de manutenção. Através das interfaces padronizadas dos módulos, algoritmos de diagnóstico, para a identificação de diferentes tipos de falhas, e de prognóstico, previsão das condições futuras de funcionamento, por exemplo, poderiam ser incluídos sem praticamente nenhuma mudança nos algoritmos e módulos já presentes no sistema.

3 PROCESSAMENTO DE SINAIS PARA DETECÇÃO E CORREÇÃO DE DRIFT

Pode-se dividir o processo de validação de sensores em quatro etapas principais: pré-processamento, processamento, pós-processamento dos sinais e o monitoramento de condição. Cada uma dessas etapas é detalhada a seguir.

3.1 PRÉ-PROCESSAMENTO DOS SINAIS

A etapa de pré-processamento dos sinais consiste na preparação dos dados que serão utilizados pelos modelos empíricos, tanto para o treinamento quanto para a predição. O condicionamento dos dados às técnicas de modelagem normalmente é necessário para que os modelos consigam extrair dos sinais as informações corretas e realmente relevantes para futuras predições.

Pode-se dividir a etapa de pré-processamento em quatro tarefas: limpeza dos dados, normalização, seleção das variáveis de entrada do modelo e seleção das amostras de treinamento.

3.1.1 Limpeza dos Dados

A limpeza dos dados consiste na filtragem e remoção de dados espúrios.

A filtragem dos sinais tem por objetivo retirar ou reduzir os ruídos normalmente presentes nas medições dos sensores. “O uso de sinais filtrados tende a facilitar o processo de treinamento dos modelos, uma vez que se reduz a complexidade dos sinais a serem aprendidos pelo modelo” (Mauro Vitor de Oliveira, 2005). Entretanto, é importante garantir que o processo de filtragem não remova informações “verdadeiras” sobre a variável mensurada, o que poderia comprometer o desempenho dos modelos.

Dados espúrios ou *outliers* são dados inconsistentes com a maioria das medidas coletadas. Medidas espúrias são normalmente causadas por erros grosseiros na medição, erros no armazenamento, e/ou outros eventos anormais. Tais dados não representativos podem afetar seriamente os modelos (CHERKASSKY; MULIER, 2007) e devem ser corrigidos ou simplesmente removidos do conjunto de dados utilizados para treinamento.

3.1.2 Normalização

A normalização dos dados consiste em tornar igual a escala de valores dos sinais de todos os sensores envolvidos. Diferentes variáveis mensuradas pelos sensores naturalmente possuem diferentes escalas, ou seja, suas próprias unidades de medida. Para algumas técnicas de modelagem, a escala das variáveis não representa um problema, entretanto, outros métodos, principalmente os baseados em distância, são sensíveis às escalas das variáveis de entrada. Nestes métodos, variáveis caracterizando pressão, por exemplo, possuiriam maior influência quando expressadas em Pascal que em kgf/m^2 .

Para técnicas de aprendizagem de máquina, uma das formas comuns de se normalizar sinais é escalá-los no hipercubo $[0, 1]^p$, onde p é o número de variáveis de entrada do modelo. Neste caso, a normalização de uma amostra $r_p \in \mathbb{R}$ obtida de um sensor p é dada pela Eq. (3.1), onde $r_{norm,p}$ é o valor normalizado de r_p , $r_{max,p}$ e $r_{min,p}$ são os valores máximos e mínimos dos dados de treinamento relativos ao sensor p . No caso de sistemas OLM, as novas amostras dos sensores, coletadas ao longo do tempo, são normalizadas de acordo com os valores de referência ($r_{max,p}$ e $r_{min,p}$) dos dados de treinamento dos modelos. Na ocorrência de novas amostras com valores maiores que $r_{max,p}$ ou menores que $r_{min,p}$, seus valores normalizados serão maiores que 1 ou menores que 0, respectivamente.

$$r_{norm,p} = \frac{r_p - r_{min,p}}{r_{max,p} - r_{min,p}} \quad (3.1)$$

3.1.3 Seleção das Variáveis de Entrada

Este passo consiste na seleção de quais variáveis, ou sensores, serão incorporados no modelo, com o objetivo de agrupar aquelas que possuem certo grau de correlação. Como os modelos empíricos baseiam suas previsões nas correlações das variáveis de entrada, o agrupamento é uma tarefa de fundamental importância. A adição de variáveis não relevantes (sem correlação com a variável a ser predita) tende a gerar modelos instáveis, cujas previsões apresentam alta variância. Entretanto, para algumas técnicas de regressão, como a regressão linear pela solução da equação normal (GRIBOK; HINES; UHRIG, 2000), variáveis altamente correlacionadas podem causar singularidade na matriz de dados entrada e, conseqüentemente, previsões instáveis. Além disso, quanto maior o número de variáveis de entrada de um modelo empírico de regressão, maior deve ser o número de amostras de treinamento para que o modelo cubra o mínimo possível do espaço de entrada.

Para os casos de redundância de *hardware* com técnicas de regressão que suportam colinearidades, o processo de seleção é simples, uma vez que apenas sensores redundantes são incluídos em um modelo. Para os casos onde redundâncias não necessariamente existem, o conjunto de sensores que se deseja monitorar deve ser separado em grupos menores altamente correlacionados. Segundo Hines e Seibert (HINES; SEIBERT, 2006), agrupamentos ótimos normalmente contém menos de 30 sensores e a adição de sinais irrelevantes acaba por aumentar a variância das previsões de um modelo, enquanto a ausência de sinais relevantes tende a causar polarização.

Um indicador de agrupamento normalmente utilizado é o coeficiente de correlação (covariância entre os sinais dividida pelo produto dos desvios padrão de cada um) para cada par de sensores. Se o coeficiente de correlação ultrapassa um determinado valor, então o correspondente par de sensores deve participar do mesmo grupo. Entretanto, este método não garante a otimalidade de agrupamento, pois variáveis que são fortemente correlacionadas fisicamente podem apresentar baixos coeficientes de correlação devido a ilusões estatísticas de conjunto de dados polarizados. Normalmente, esse método é utilizado acompanhado de senso de engenharia, ou seja, selecionam-se as variáveis considerando também conhecimentos especialistas sobre a dinâmica do processo.

Técnicas de exploração do espaço de possibilidades também podem ser empregadas para o agrupamento das variáveis, como algoritmos genéticos (Mauro Vitor de Oliveira, 2005) e a *stepwise grouping* (AN; HEO; CHANG, 2011). Basicamente, essas técnicas tentam minimizar o número de tentativas possíveis para se alcançar um solução ótima ou subótima de agrupamento. Cada tentativa implica na criação de um diferente modelo e na avaliação da qualidade das previsões geradas por ele. Através de uma heurística, reduz-se o espaço de possibilidades de combinação das variáveis para, finalmente, selecionar aquele agrupamento que gera o modelo com melhor qualidade de predição.

3.1.4 Seleção das Amostras de Treinamento

Dois aspectos principais, intrinsecamente relacionados, devem ser considerados para a seleção das amostras para o treinamento dos modelos: quantidade e representatividade.

Normalmente, quanto maior a quantidade de amostras que se tem de um sistema, mais informações ficam disponíveis e mais preciso é o modelo que se pode obter para representá-lo. Entretanto, os limites físicos de armazenamento de dados e a capacidade computacional para processá-los devem

ser considerados ao se implementar um sistema OLM. Quantidades excessivas de dados podem custar muito tempo para treinamento dos modelos e para os cálculos das predições, a um ganho insignificante de informações sobre o sistema.

A representatividade está relacionada justamente à quantidade de informações relevantes sobre um sistema presentes em um conjunto de dados. Normalmente, a quantidade de amostras de um conjunto de dados não é linearmente relacionado à representatividade deste. Em muitos casos é possível obter pequenos conjuntos de dados que possuem praticamente a mesma representatividade que conjuntos volumosos (HINES et al., 2008b).

Portanto, a seleção das amostras de treinamento de um modelo regressivo deve considerar primeiramente a cobertura de todos os futuros pontos de operação da planta ou poço para, finalmente, se fazer uma análise de ganho de precisão das predições em relação à quantidade de amostras. Além disso, é de suma importância certificar que as amostras não correspondem a pontos de operação atípicos, como em casos de falha em equipamento ou de desvios em sensores. Se presentes no conjunto de treinamento do modelo, os pontos de operação atípicos são aprendidos pelo modelo como comportamento normal dos sensores.

Em sistemas dinâmicos como poços de petróleo, onde não se tem dados de histórico que abranja todos os possíveis pontos de operação futuros devido às mudanças das características do próprio poço, normalmente torna-se necessário o retreinamento do modelo regressivo com novas amostras ao longo da vida útil do poço.

3.2 PROCESSAMENTO DOS SINAIS

O processamento dos sinais é a etapa de predição propriamente dita. A partir de novas amostras das variáveis de entrada, o modelo empírico gera predições sobre a variável de saída. Os principais aspectos a serem considerados nesta etapa são: a escolha do tipo de estrutura do modelo, a escolha da técnica de regressão e a construção/otimização do modelo.

3.2.1 Seleção do Tipo de Estrutura

Normalmente, a estrutura dos modelos pode ser classificada em três diferentes tipos: inferencial, hetero-associativa ou auto-associativa.

Um modelo inferencial utiliza um conjunto de variáveis de entrada $[r_1, \dots, r_n]$ para inferir o valor de uma única variável de saída y . Esse modelo

pode ser expandido para um estrutura hetero-associativa, onde um conjunto de variáveis de entrada $[r_1, \dots, r_p]$ é usado para prever os valores de um conjunto de variáveis de saída $[y_1, \dots, y_q]$.

Um modelo auto-associativo é normalmente treinado para emular as próprias variáveis de entrada, ou seja, um conjunto de variáveis $[r_1, \dots, r_p]$ é usado para gerar previsões $[\hat{r}_1, \dots, \hat{r}_p]$, que correspondem aos valores “corrigidos” das variáveis de entrada de acordo com o conjunto de dados de treinamento usado para a construção do modelo. Dessa forma, se as variáveis de entrada possuem a mesma correlação existente nos dados de treinamento, os valores de saída do modelo auto-associativo tendem a ser iguais aos valores de entrada. Pequenas alterações nessas correlações, como as causadas por algumas falhas em sensores, tendem a ser corrigidas pela correlação capturada pelo modelo através dos dados de treinamento.

3.2.2 Técnicas de Regressão

As duas técnicas de regressão abordadas neste trabalho são: AAKR e SVM.

3.2.2.1 AAKR

A regressão por kernel auto-associativa é uma técnica não paramétrica para modelagem baseada em histórico de dados, cujas previsões se baseiam na similaridade entre as entradas do modelo e um histórico de dados armazenado. A arquitetura do modelo AAKR apresentada a seguir é uma variação da regressão inferencial multivariável por kernel, apresentada por Wand e Jones (WAND; JONES, 1995).

Considerando \mathbf{X} uma matriz de histórico de amostras livres de erros, ou memória, $x_j^{(i)}$ representa a i -ésima observação da j -ésima variável/sensor. Para n_m vetores de memória e p variáveis, a matriz \mathbf{X} é definida como

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n_m)} & x_2^{(n_m)} & \dots & x_p^{(n_m)} \end{bmatrix}.$$

As entradas do modelo são definidas pelo vetor de amostras \mathbf{r} , definido por

$$\mathbf{r} = [r_1 \quad r_2 \quad \cdots \quad r_p] .$$

A predição do valor corrigido de \mathbf{r} é calculada como uma média ponderada das observações livres de erros contidas em \mathbf{X} . O funcionamento do modelo AAKR é composto de três passos básicos. Primeiro, calcula-se as distâncias entre \mathbf{r} e cada linha da matriz de memória \mathbf{X} . Dentre as possíveis métricas de distância, a distância Euclidiana é a mais comumente utilizada (HINES et al., 2008a), dada pela Eq. (3.2), onde u_i é a distancia Euclidiana entre o vetor de memória $\mathbf{x}^{(i)}$ e a entrada \mathbf{r} .

$$u_i(\mathbf{X}^{(i)}, \mathbf{r}) = \sqrt{(x_1^{(i)} - r_1)^2 + (x_2^{(i)} - r_2)^2 + \cdots + (x_p^{(i)} - r_p)^2} \quad (3.2)$$

Este cálculo é repetido para todos os n_m vetores de memória da matriz \mathbf{X} , resultando em um vetor coluna de distâncias

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n_m} \end{bmatrix} .$$

Em seguida, as distâncias são convertidas em medidas de similaridade ou pesos através da função kernel, como a Gaussiana apresentada na Eq. (3.3), onde \bar{w}_i é o i -ésimo peso relativo a i -ésima distância u_i e h é a largura de banda da função kernel K . Apesar de outras funções kernel possuírem vantagens e desvantagens em certas situações, segundo trabalhos indicados por Hines et al. (HINES et al., 2008a) e An, Heo e Chang (AN; HEO; CHANG, 2011), a função kernel não desempenha um papel crítico no desempenho de modelos localmente ponderados, como o AAKR.

$$\bar{w}_i = K(u_i) = \frac{1}{\sqrt{2\pi}h^2} \exp\left(-\frac{u_i^2}{h^2}\right), \quad (3.3)$$

Por fim, através da Eq. (3.4), os pesos são combinados com os vetores de memória para produzir uma estimativa \hat{r}_j de cada variável r_j . Calculando as estimativas para $j = \{1, 2, \dots, p\}$, obtém-se a saída do modelo, a predição $\hat{\mathbf{r}}$. Alternativamente, definindo $a = \sum_{i=1}^{n_m} \bar{w}_i$, o estimador AAKR pode ser apresentado pela Eq. (3.5).

$$\hat{f}_j = \frac{\sum_{i=1}^{n_m} \bar{w}_i \cdot x_j^{(i)}}{\sum_{i=1}^{n_m} \bar{w}_i} \quad (3.4)$$

$$\hat{\mathbf{f}} = \frac{\bar{\mathbf{w}}^T \mathbf{X}}{a} \quad (3.5)$$

3.2.2.1.1 Seleção dos vetores de memória

A seleção dos vetores de memória é uma tarefa crítica para o desenvolvimento de modelos não paramétricos como o AAKR (HINES et al., 2008b), onde as predições do modelo se baseiam na similaridade entre suas entradas e cada um dos vetores de memória. Um conjunto de dados de treinamento muito grande tornaria muito elevado o custo computacional de cada predição. O que normalmente se faz é selecionar um subconjunto dos dados de treinamento, visando diminuir o custo computacional sem perder informações significantes sobre a correlação entre as variáveis.

Basicamente, duas questões devem ser consideradas para a seleção dos vetores de memória: qual é a *quantidade* adequada e *como* realizar essa seleção. Em relação à quantidade de vetores n_m , o valor deve ser adequado às necessidades do sistema OLM, retratando um balanceamento entre o custo computacional (ou velocidade do cálculo das predições) e a qualidade das predições (HINES et al., 2008b). Modelos com alto n_m tendem a gerar melhores predições, entretanto, o custo computacional de cada uma delas pode ser impraticável, caso o tempo de uma predição seja crítico.

Em relação à forma de seleção dos vetores, dois métodos simples e comumente utilizados são a seleção por mínimos-máximos e a seleção por ordenamento dos vetores (HINES et al., 2008a). O método de seleção por mínimos-máximos é dividido em duas etapas. Primeiro, particiona-se o conjunto de dados em n_b bandas, de acordo com a Eq. (3.6), onde n_m é o número de vetores a serem selecionados e p é a quantidade de variáveis nos dados. Em seguida, para cada banda gerada, selecionam-se os vetores que contenham os valores máximos e mínimos de cada variável, sem repetição de vetores.

$$n_b = \frac{n_m}{2p} \quad (3.6)$$

No método de seleção por ordenamento de vetores, primeiramente ordenam-se os vetores de treinamento de acordo com a norma Euclidiana. A Equação (3.7) apresenta o cálculo da norma Euclidiana N_i para um vetor

$\mathbf{x}^{(i)}$ de treinamento. Em seguida, selecionam-se n_m vetores amostrando de forma periódica. É importante destacar que este método é intrinsecamente relacionado com a localização da origem do espaço da matriz de dados de treinamento, o que torna importante escalar as variáveis antes da utilização deste método de seleção.

$$N_i = \sqrt{x_1^{(i)2} + x_2^{(i)2} + \dots + x_p^{(i)2}} \quad (3.7)$$

Neste trabalho a seleção dos vetores de memória foi realizada pela combinação dos dois métodos anteriores. Como normalmente o método de seleção por mínimos-máximos não obtém o número n_m de vetores desejado (um mesmo vetor pode conter os valores máximos ou mínimos de mais de uma variável), usa-se o método de ordenamento de vetores para selecionar a quantidade de vetores necessária para se obter n_m .

3.2.2.2 SVM

Na técnica de SVM para regressão, um conjunto de dados de treinamento $\{(\mathbf{r}^{(1)}, y^{(1)}), \dots, (\mathbf{r}^{(n_m)}, y^{(n_m)})\}$, onde $\mathbf{r} \in \mathbb{R}^p$ e $y \in \mathbb{R}$, é usado para estimar uma função $f(\mathbf{r})$ que correlaciona as p variáveis de \mathbf{r} com uma variável dependente y a partir de n_m observações independentes e identicamente distribuídas. Dessa forma, um modelo SVM apresenta estrutura do tipo inferencial.

A técnica corresponde à minimização da Eq. (3.8), onde $f(\mathbf{r}) = \langle \mathbf{w}, \mathbf{r} \rangle + b$, com $w \in \mathbb{X}$ (espaço de características) e $b \in \mathbb{R}$, é o estimador da função de dependência entre \mathbf{r} e y . A solução da Eq. (3.8) pode ser escrita como a Eq. (3.9). Essa formulação torna a solução esparsa no sentido que erros menores que ε são ignorados, ou seja, a solução torna-se “ ε -insensível”.

$$\begin{aligned} \text{Minimizar} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \left(\xi^{(i)} + \xi^{*(i)} \right) \\ \text{Sujeito a} \quad & \left\{ \begin{array}{l} y^{(i)} - \langle \mathbf{w}, \mathbf{r}^{(i)} \rangle - b \leq \varepsilon + \xi^{(i)} \\ \langle \mathbf{w}, \mathbf{r}^{(i)} \rangle + b - y^{(i)} \leq \varepsilon + \xi^{*(i)} \\ \xi^{(i)}, \xi^{*(i)} \geq 0 \end{array} \right\} \end{aligned} \quad (3.8)$$

$$f(\mathbf{r}) = \sum_{j=1}^p w_j r_j + b \quad (3.9)$$

O primeiro termo da equação (3.8) é um termo de regularização que evita o mau condicionamento do problema de estimação, contribuindo na simplicidade do estimador da função. O segundo termo é a função de perda ε -insensível, a qual mede o quanto os valores estimados estão próximos dos valores y . As variáveis $\xi^{(i)}$ e $\xi^{*(i)}$ são chamadas de variáveis de folga, que determinam o grau de penalização aplicados às estimativas com erros maiores que ε . A Figura 6 ilustra a zona de insensibilidade formada por ε e a penalização efetuada pelas variáveis de folga. Por isso, qualquer erro absoluto menor que ε não necessita de valores diferentes de zero para as variáveis $\xi^{(i)}, \xi^{*(i)}$ na função objetivo, o que causa a esparsividade da solução. A constante C determina o custo-benefício entre a complexidade da função f e a quantidade tolerada de desvios maiores que ε .

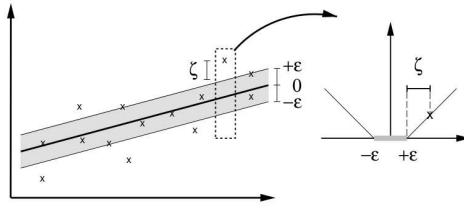


Figura 6 – Zona de insensibilidade e penalização no modelo SVM (SMOLA; SCHÖLKOPF, 2004).

Escrevendo a Eq. (3.8) na forma dual e resolvendo por diferenciação em relação às variáveis primais, o problema resulta em maximizar a função $W(\alpha^*, \alpha)$ na Eq. (3.10), onde α, α^* são multiplicadores de Lagrange e $K(\mathbf{r}^{(i)}, \mathbf{r}^{(j)})$ é a função kernel que substitui o que seriam os produtos internos $\langle \mathbf{r}^{(i)}, \mathbf{r}^{(j)} \rangle$ e permite lidar com não linearidades do espaço de entrada do modelo.

Maximizar

$$W(\alpha^*, \alpha) = -\varepsilon \sum_{i=1}^{n_m} (\alpha^{(i)} + \alpha^{*(i)}) + \sum_{i=1}^{n_m} y^{(i)} (\alpha^{(i)} - \alpha^{*(i)}) - \frac{1}{2} \sum_{i,j=1}^{n_m} (\alpha^{(i)} - \alpha^{*(i)}) (\alpha_j - \alpha_j^*) K(\mathbf{r}^{(i)}, \mathbf{r}^{(j)}) \quad (3.10)$$

Sujeito a
$$\left\{ \begin{array}{l} \sum_{i=1}^{n_m} (\alpha^{*(i)} - \alpha^{(i)}) = 0 \\ 0 \leq \alpha^{(i)}, \alpha^{*(i)} \leq C \end{array} \right\}$$

A função $f(\mathbf{r})$ passa a ser dada pela Eq. (3.11). Os vetores $\{\mathbf{r}^{(1)}, \dots, \mathbf{r}^{(N_{sv})}\}$, pertencentes ao conjunto inicial de treinamento, são chamados de “vetores suporte”, onde N_{sv} (normalmente $N_{sv} \ll n_m$) é o número de vetores suporte correspondentes à quantidade de valores de $\{y^{(1)}, \dots, y^{(n_m)}\}$ que estão pelo menos ε afastados de suas estimativas $\{f(\mathbf{r}^{(1)}), \dots, f(\mathbf{r}^{(n_m)})\}$.

$$f(\mathbf{r}) = \sum_{i=1}^{N_{sv}} (\alpha^{*(i)} - \alpha^{(i)}) \mathcal{K}(\mathbf{r}, \mathbf{r}^{(i)}) + b \quad (3.11)$$

As condições de Kuhn-Tucker demandam que o produto entre as variáveis duais e as restrições devem desaparecer para otimalidade. Por isso, para $|f(\mathbf{r}^{(i)}) - y^{(i)}| \geq \varepsilon$, os multiplicadores de Lagrange são diferentes de zero; e para pontos dentro da região ε -insensível, os coeficientes $\alpha^{*(i)}, \alpha^{(i)}$ devem desaparecer. Os vetores com coeficientes diferentes de zero são os chamados vetores suporte. Basicamente, os vetores suporte são os vetores que suportam a “superfície de decisão” ou o “hiperplano” que melhor se adequa aos dados, de acordo com o critério especificado na Eq. (3.8).

A configuração do modelo SVM consiste na escolha da função kernel e de seus parâmetros, e na especificação de ε e C . Usualmente, tais parâmetros são selecionados por *cross-validation* (SMOLA; SCHÖLKOPF, 2004), entretanto, neste trabalho apenas um tipo de função kernel foi utilizada, a função de base radial (RBF). Apresentada na Eq. (3.12), onde γ é um parâmetro de configuração de suavidade, a função RBF é normalmente uma escolha razoável quando pouco se sabe sobre os dados (SMOLA; SCHÖLKOPF, 2004).

$$\mathcal{K}(\mathbf{r}, \mathbf{r}^{(i)}) = \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}^{(i)}\|^2}{2\gamma^2}\right) \quad (3.12)$$

A constante C determina o custo-benefício entre a complexidade do modelo (suavidade) e o grau de tolerância dos desvios maiores que ε em (3.8) (CHERKASSKY; MA, 2004). Valores muito grandes de C tornam o problema sem restrição; valores pequenos atribuem mais peso para a regularização. O parâmetro ε controla a largura da zona de insensibilidade, usada para adequar os dados de treinamento. O valor de ε é o parâmetro que possui maior efeito sobre o número de vetores suporte. Quanto maior for ε , menos vetores suporte são utilizados e mais “suave” torna-se a função f .

3.2.3 Construção e Otimização dos Modelos

Partindo-se de uma técnica de regressão, constrói-se um modelo empírico a partir de um conjunto de amostras de treinamento, compostas por pares entrada/saída. As amostras de treinamento são exemplares de valores das variáveis de entrada juntamente com os valores das variáveis de saída que se espera do modelo. No caso de um modelo SVM, as amostras são compostas pelos valores de p variáveis de entrada e um valor correspondente à variável de saída. Para um modelo AAKR, os pares entrada/saída possuem o mesmo valor.

Para ambas as técnicas, a complexidade dos modelos podem ser especificadas por parâmetros e deve ser ajustada para que se adeque à complexidade dos dados de treinamento. Se o modelo não é flexível, ele não será capaz de modelar as relações dos dados de treinamento; se o modelo é excessivamente complexo, problemas de sobreajustamento (*overfitting*) dos dados podem ocorrer, incluindo ruídos na modelagem. Considerando o uso de funções kernel do tipo Gaussiana ou RBF, a complexidade dos modelos (o grau de ajuste do modelo aos dados de treinamento) SVM é determinada pelos parâmetros C , ϵ e γ ; nos modelos AAKR, pelo parâmetro h . Em muitas aplicações, esses valores podem ser encontrados por tentativa e erro, ou *grid-search* com *cross-validation* (HSU et al., 2003), onde treinam-se vários modelos com diferentes permutações de valores de configuração e avalia-se a qualidade das predições destes através de um novo conjunto de amostras. O modelo com a configuração de melhor desempenho é adotado como o modelo final para o sistema de validação.

Durante a construção de um modelo regressivo, um método simplório utilizado para se determinar a complexidade adequada do modelo ao problema de aplicação consiste nos seguintes passos:

1. obter dois diferentes conjuntos de dados além do conjunto de treinamento, que podem ser chamados de conjunto de otimização e conjunto de teste;
2. calcular o erro de predição de modelos de diferentes complexidades para o conjunto de otimização;
3. selecionar o modelo que tenha apresentado o menor erro de predição do conjunto de dados de otimização;
4. confirmar o desempenho do modelo selecionado calculando o erro de predição para o conjunto de teste.

Os conjuntos de dados de otimização e teste devem ser semelhantes ao conjunto de treinamento, no sentido de corresponder a comportamentos semelhantes do sistema, como pontos de operação de uma planta. O erro de predição é calculado a partir do resultado $f(\mathbf{r})$ do modelo e a saída real y do sistema para uma dada entrada \mathbf{r} . Geralmente se usa o erro quadrático médio, como apresentado pela Eq. (3.13), como representante do erro de predição, onde n é a quantidade de observações utilizadas para a construção do modelo.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - f(\mathbf{r}^{(i)}))^2 \quad (3.13)$$

Segundo Cherkassky e Mulier (CHERKASSKY; MULIER, 2007), existe o seguinte consenso: para métodos flexíveis de aprendizado com um número finito de amostras (como os métodos baseados em kernel utilizados neste trabalho), as predições com melhor desempenho são obtidas por modelos de complexidade ótima. Deve-se preferir modelos mais simples a modelos complexos e otimizar a relação entre complexidade e acurácia do modelo para o conjunto de dados de treinamento.

Existem métodos mais completos para o desenvolvimento e avaliação de modelos regressivos¹, mas o método descrito anteriormente é suficiente para a proposta deste trabalho.

3.3 PÓS-PROCESSAMENTO DOS SINAIS

A etapa de pós-processamento dos sinais consiste na estimação dos desvios entre as predições dos modelos empíricos e as variáveis que se deseja estimar. A partir das estimativas dos desvios, é possível realizar uma “pré-correção” das leituras provenientes dos sensores que servem como entrada dos modelos empíricos, permitindo que as predições fiquem mais próximas dos verdadeiros valores das variáveis.

A seguir, apresenta-se a construção de um filtro de Kalman para a correção de sensores que apresentem *drift*.

¹<http://www.cedar.buffalo.edu/~srihari/CSE574/Chap3/Bias-Variance.pdf>
<http://blog.stephenpurpura.com/post/13052575854/managing-bias-variance-tradeoff-in-machine-learning>
<http://www.cedar.buffalo.edu/~srihari/CSE555/Chap9.Part2.pdf>

3.3.1 Filtro de Kalman

Um possível modelo matemático usado para estimar a amplitude do *drift* d de um sensor é apresentado na Eq. (3.14), onde $v^{(k)}$ é considerado um ruído Gaussiano de média zero e variância σ_v .

$$d^{(k)} = d^{(k-1)} + v^{(k)}, \quad v^{(k)} \sim \mathcal{N}(0, \sigma_v) \quad (3.14)$$

Em rastreamento de alvos (*target tracking*), a Equação (3.14) é conhecida como o modelo matemático do comportamento dinâmico do alvo (TAKRURI; RAJASEGARAR; CHALLA, 2008). No caso de detecção de *drift*, o objetivo é rastrear ou acompanhar a amplitude do *drift* de um sensor. Se for assumido que um sensor apresenta desvios de forma suave, vagarosamente crescente, linear ou exponencial, pode-se considerar o modelo da Eq. (3.14) como uma aproximação razoável. Uma outra consideração deste modelo é a decorrelação entre *drifts* de diferentes sensores, mesmo sendo as variáveis de processo correlacionadas.

A observação do desvio no sensor estimado, dado por

$$z = r - \hat{r},$$

não corresponde ao valor verdadeiro do desvio, uma vez que este valor não está disponível. Então, pode-se descrever a equação de medição como a Eq. (3.15), onde $q^{(k)}$ é um ruído Gaussiano de variância σ_q .

$$z = d^{(k)} + q^{(k)}, \quad q^{(k)} \sim \mathcal{N}(0, \sigma_q) \quad (3.15)$$

Quando uma nova observação r e a estimacão de seu valor corrigido \hat{r} estiverem disponíveis, os seguintes passos são executados para se obter uma nova estimacão do desvio $d^{(k)}$ eventualmente presente no sensor:

1. predicção,

$$\hat{d}^{(k|k-1)} = \hat{d}^{(k-1)}$$

2. MSE mínimo da predicção,

$$M^{(k|k-1)} = M^{(k-1)} + \sigma_v,$$

3. ganho de Kalman,

$$K = \frac{M^{(k|k-1)}}{\sigma_q + M^{(k|k-1)}}$$

4. correção da estimativa,

$$\hat{d}^{(k|k)} = \hat{d}^{(k|k-1)} + K(z - \hat{d}^{(k|k-1)})$$

5. mínimo MSE,

$$M^{(k|k)} = (1 - K)M^{(k|k-1)}.$$

Para o emprego deste modelo de filtro de Kalman na estimação de desvios em sensores é necessário determinar os valores dos parâmetros σ_v e σ_q para cada uma das variáveis monitoradas. No trabalho de Takruri (TAKRURI; RAJASEGARAR; CHALLA, 2008), os parâmetros são escolhidos por tentativa e erro. Se um valor alto é usado para σ_q , as estimativas do desvio demoram mais para seguir o desvio real; para valores pequenos de σ_q , as estimativas seguem qualquer pequena diferença entre as predições \hat{r} e as leituras r do sensor. O valor de σ_v possui efeito contrário: valores altos implicam em estimativas instáveis do desvio que seguem qualquer pequena diferença entre as predições \hat{r} e as leituras r do sensor; valores baixos implicam em maior lentidão para as estimativas seguirem o desvio real.

As estimativas do desvio de um sensor podem ser usadas pelo menos de duas formas: indicar a ocorrência de desvios intoleráveis e a consequente necessidade de manutenção do sensor; e corrigir as leituras do sensor antes mesmo de serem utilizadas para predição pelo modelo empírico, o que teoricamente proporcionaria predições mais próximas do valor verdadeiro da variável mensurada.

3.4 MONITORAMENTO DE CONDIÇÃO

A etapa de monitoramento da condição visa monitorar a condição de funcionamento dos sensores, detectando a ocorrência de eventos e mudanças de estado.

A detecção de anomalias nos sensores é baseada na análise do erro entre as predições do modelo e as medidas produzidas por eles, ou seja, o resíduo d calculado pela Eq. (3.16), onde y é representa uma medida realizada pelo sensor que se deseja monitorar e $f(\mathbf{r})$ é uma estimativa dessa medida a partir das medidas \mathbf{r} de sensores correlacionados.

$$d = y - f(\mathbf{r}) \quad (3.16)$$

Considerando o sistema sensoreado operando na região de treinamento do modelo, as predições são virtualmente iguais às medidas realizadas pelos sensores quando estes se encontram funcionando corretamente, e os resíduos

tendem a apresentar média zero e variância semelhante a do sensor. Anomalias nas medições dos sensores, como *drift*, *outliers*, mudanças abruptas, aumento da variância (errático) etc., causam mudanças nas características estatísticas esperadas nos resíduos e normalmente podem ser detectadas por técnicas estatísticas de detecção (Mauro Vitor de Oliveira, 2005). Dentre essas técnicas, podem-se citar: verificação de limites das propriedades estatísticas, como média e a variância; auto-correlação dos resíduos; densidade de potência dos resíduos; e o método SPRT.

No caso da detecção específica de *drifts*, a verificação da incerteza das previsões do modelo também pode ser empregada (HINES; GARVEY, 2008). Os métodos analíticos para os cálculos de incerteza são específicos para cada técnica de modelagem empírica, enquanto métodos baseados em simulações Monte Carlo são gerais, mas podem ser computacionalmente custosos.

Pela simplicidade de implementação e abrangência de anomalias detectáveis, neste trabalho emprega-se o método SPRT, descrito a seguir.

3.4.1 SPRT

O método conhecido por teste de razão de probabilidade sequencial (SPRT), desenvolvido por Wald (WALD, 1947), consiste em testes estatísticos de hipóteses e tem sido aplicado frequentemente para a detecção de anomalias em sistemas. No caso de detecção de falhas em sensores, duas hipóteses são testadas a cada nova amostra de dados do sistema: modo normal de operação, ou H_0 , e modo degradado de operação, ou H_1 . Aplicado aos resíduos gerados pela Eq. (3.16), para cada nova amostra $d^{(i)}$ do resíduo gerado para o sensor monitorado, uma das duas possíveis hipóteses é aceita através da execução do seguinte procedimento:

1. cálculo do logaritmo da razão de verossimilhança, dado pela Eq. (3.17);

$$\Lambda^{(i)} = \ln \frac{P(d^{(i)}|H_1)}{P(d^{(i)}|H_0)} \quad (3.17)$$

2. cálculo da soma acumulativa de Λ , dado pela Eq. (3.18);

$$S^{(i)} = S^{(i-1)} + \Lambda^{(i)} \quad (3.18)$$

3. verificação da regra de parada, um esquema simples de verificação de limites:

- se $\ln A < S^{(i)} < \ln B$, continua-se o monitoramento e o cálculo de

$$S^{(i+1)} = S^{(i)} + \Lambda^{(i+1)} \text{ (passo 2);}$$

- se $S_p^{(i)} \leq \ln A$, H_0 é aceita;
- se $S_p^{(i)} \geq \ln B$, H_1 é aceita e um aviso ou alarme é emitido.

$P(d^{(i)}|H_1)$ é a probabilidade de observar $d^{(i)}$ dado que H_1 é verdadeira. A e B são respectivamente o limite inferior e superior, cujos valores podem ser definidos pela probabilidade de ocorrência de falsos alarmes, α , e pela probabilidade de ocorrência de alarmes perdidos, β , como

$$A = \frac{\beta}{1 - \alpha}, \quad B = \frac{1 - \beta}{\alpha}.$$

Supondo os resíduos como Gaussianos, normalmente o modo H_0 corresponde a um sinal com média zero e variância semelhante a do sensor em questão, enquanto o modo H_1 corresponde a uma média diferente de zero ou a uma mudança no valor da variância. Neste trabalho, assume-se que o *drift* corresponda a uma mudança $\pm M$ na média dos resíduos, enquanto a variância σ^2 se mantém inalterada. Portanto, a função de distribuição de probabilidade dos resíduos de um sensor para o modo H_0 é dada pela Eq. (3.19).

$$P(d^{(i)}|H_0) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-d^{(i)2}}{2\sigma^2}\right) \quad (3.19)$$

Considerando o modo degradado H_1 representado por mudanças positivas ($+M$) ou negativas ($-M$) na média, o logaritmo da razão de verossimilhança é dado pela Eq. (3.20). Como sugerido por Hines e Garvey (HINES; GARVEY, 2008), o valor de M pode ser determinado numericamente aplicando o SPRT em conjuntos de dados livres de falhas e localizando um valor de M que resulte em uma probabilidade de alarmes falsos semelhante a α .

$$\Lambda^i = \begin{cases} M/\sigma^2(d^i - M/2), & \text{para uma mudança positiva} \\ M/\sigma^2(-d^i - M/2), & \text{para uma mudança negativa} \end{cases} \quad (3.20)$$

4 APLICAÇÃO DOS SISTEMAS IMPLEMENTADOS A DADOS DE POÇOS DE PETRÓLEO

Neste capítulo são apresentados os sistemas implementados para monitoramento e validação de sensores de poços de petróleo. Os sistemas implementados são avaliados com diferentes conjuntos de dados, gerados a partir de simulações ou a partir de sensores reais. As métricas de desempenho e os resultados das avaliações também são apresentados a seguir.

4.1 SISTEMAS DE VALIDAÇÃO DE SENSORES

Descrevem-se a seguir três diferentes sistemas de validação de sensores implementados neste trabalho, os quais se distinguem pela técnica de modelagem e/ou pelo emprego do filtro de Kalman.

4.1.1 Sistema 1 — AAKR-SPRT

O Sistema 1 é semelhante a um dos sistemas utilizados por Hines et al. (HINES et al., 2008a), onde a técnica de modelagem utilizada é a AAKR e a detecção de desvios é realizada pelo algoritmo SPRT. Na Figura 7 é ilustrado o esquema de funcionamento do Sistema 1.

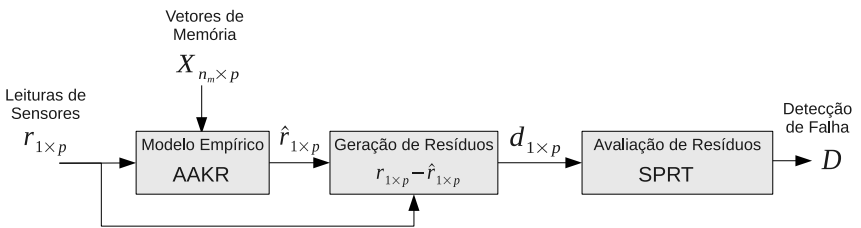


Figura 7 – Diagrama do processo de monitoramento pelo Sistema 1.

O modelo regressivo AAKR é desenvolvido usando o conjunto de dados $\mathbf{X}_{n_m \times p}$, que permanece armazenado como memória para geração de previsões. Quando um novo vetor de leituras dos sensores $\mathbf{r}_{1 \times p}$ está disponível, uma nova previsão $\hat{\mathbf{r}}_{1 \times p}$ é calculada. A diferença entre a leitura $\mathbf{r}_{1 \times p}$ e sua previsão $\hat{\mathbf{r}}_{1 \times p}$, o resíduo $\mathbf{d}_{1 \times p}$, é analisada pelo algoritmo estatístico SPRT para a indicação de ocorrências de falhas ou desvios, $D \in [0, 1]$.

4.1.2 Sistema 2 — SVM-SPRT

O Sistema 2 é composto por um modelo de regressão SVM e pelo algoritmo estatístico SPRT. Como descrito na Seção 3.2.2, um modelo SVM possui estrutura inferencial, sendo capaz de gerar predições para apenas uma variável. Por consequência, para gerar predições de p variáveis, como o modelo AAKR, são necessários p diferentes modelos SVM. Na Figura 8 é apresentado o esquema de funcionamento do Sistema 2 para o monitoramento de uma variável.

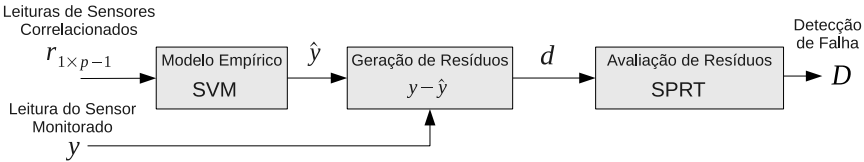


Figura 8 – Diagrama do processo de monitoramento de um único sensor pelo Sistema 2.

Considerando o monitoramento de uma variável dentre um conjunto de p variáveis correlacionadas, quando novas leituras $\mathbf{r}_{1 \times p-1}$ e y (onde y corresponde à leitura do sensor monitorado) estão disponíveis, uma nova predição \hat{y} é calculada pelo modelo SVM. A diferença entre a leitura y do sensor monitorado e sua predição \hat{y} , o resíduo d , é analisada pelo algoritmo estatístico SPRT para a indicação de ocorrência de falhas ou desvios, $D \in [0, 1]$.

4.1.3 Sistema 3 — AAKR-KF-SPRT

O Sistema 3 é composto por um modelo de regressão AAKR, um filtro de Kalman e o algoritmo estatístico SPRT. Na Figura 9 é ilustrado o esquema de funcionamento do Sistema 3.

No Sistema 3, as entradas do modelo AAKR correspondem às leituras corrigidas $\mathbf{x}_{1 \times p}$ dos p sensores a partir de uma estimativa inicial dos desvios $\hat{\mathbf{d}}_{1 \times p}^{k-1}$ e das leituras reais $\mathbf{r}_{1 \times p}$. As predições $\hat{\mathbf{x}}_{1 \times p}$ geradas pelo modelo AAKR e as leituras reais $\mathbf{r}_{1 \times p}$ são usadas como entradas do filtro de Kalman para a atualização das estimativas dos desvios, $\hat{\mathbf{d}}_{1 \times p}^k$. Finalmente, as novas estimativas $\hat{\mathbf{d}}_{1 \times p}^k$ são analisadas pelo algoritmo SPRT para a indicação de ocorrência de falhas ou desvios, $D \in [0, 1]$.

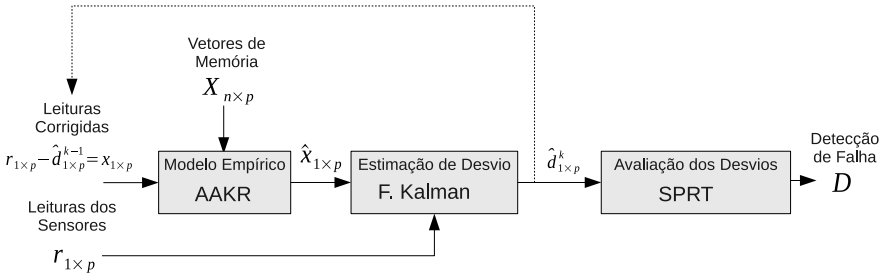


Figura 9 – Diagrama do processo de monitoramento pelo Sistema 3.

4.2 MÉTRICAS DE DESEMPENHO

Segundo Hines e Garvey (HINES; GARVEY, 2008), tradicionalmente o desempenho de sistemas de monitoramento de calibração de sensores é mensurada a partir de três indicadores: acurácia, autosensibilidade e sensibilidade cruzada. A acurácia mensura a habilidade do modelo para gerar previsões corretas e precisas das leituras dos sensores. A autosensibilidade indica a habilidade do modelo para gerar previsões corretas de um determinado sensor quando as leituras deste estão incorretas devido a algum tipo de falha. Já a sensibilidade cruzada mensura o efeito dos dados de um sensor defeituoso sobre as previsões dos demais sensores.

A capacidade do modelo de prever precisamente e corretamente os valores do sensor é medido pelo erro de acurácia (E_a) que é normalmente apresentada como o erro médio quadrático (MSE) entre as previsões e as medições dos sensores. A Equação (4.1) apresenta o cálculo de E_a para uma variável, onde N é o número de amostras, $r^{(i)}$ é a amostra i de um sensor livre de falhas e $\hat{r}^{(i)}$ é a respectiva previsão.

$$E_a = \frac{1}{N} \sum_{i=1}^N \left(\hat{r}^{(i)} - r^{(i)} \right)^2 \quad (4.1)$$

A autosensibilidade (S_A) mede a habilidade do modelo prever corretamente quando a própria medida do sensor está incorreta em razão de alguma falha (esta métrica não é aplicada a modelos inferenciais como o SVM). A Equação (4.2) apresenta o cálculo de S_a para um determinado sensor k , onde $\hat{r}_{drift}^{(i)}$ é a previsão com desvio, $\hat{r}^{(i)}$ é a previsão sem desvio, enquanto $r_{drift}^{(i)}$ e $r^{(i)}$ são as entradas com e sem desvio, respectivamente.

$$S_{A_k} = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{r}_{k,drift}^{(i)} - \hat{r}_k^{(i)}|}{|r_{k,drift}^{(i)} - r_k^{(i)}|} \quad (4.2)$$

O efeito de uma entrada de sensor falho na predição das outras variáveis é calculado pela sensibilidade cruzada (S_C). Para um sensor sem desvio j e um sensor com desvio k , a sensibilidade cruzada é calculada como a Eq. (4.3).

$$S_{C_k} = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{r}_{j,drift}^{(i)} - \hat{r}_j^{(i)}|}{|r_{k,drift}^{(i)} - r_k^{(i)}|} \quad (4.3)$$

Além das métricas tradicionais, adota-se neste trabalho uma métrica relativa ao erro de predição (E_p) para entradas do modelo que podem conter desvios, calculado também pelo MSE. Diferentemente de E_a , onde as entradas não devem conter falhas ou desvios, o valor de E_p mede o erro das predições na presença de desvios em uma ou mais entradas do modelo.

Os melhores sistemas para predição dos sensores de poços são aqueles com baixos valores de E_a , S_A , S_C e E_p .

4.3 ENSAIOS

Os ensaios apresentados a seguir visam avaliar a aplicabilidade dos sistemas desenvolvidos a sensores de poços de petróleo. A aplicabilidade é avaliada em dois aspectos principais:

- consistência e qualidade das predições, ou seja, quão próximas estão as predições e os valores corretos/verdadeiros que os sensores deveriam gerar;
- capacidade de detecção de desvios, ou seja, os sistemas devem detectar a ocorrência de desvios nas leituras de sensores e identificar qual ou quais deles estão defeituosos.

Os ensaios consistem em utilizar os sistemas implementados a conjuntos de dados e simular situações em que se esperam os benefícios do uso de tais sistemas. Os conjuntos de dados são compostos por amostras de diferentes sensores correlacionados, localizados em diferentes pontos de um poço. Inserindo desvios em algum dos sensores, os sistemas devem ser capazes de detectar a ocorrência da falha, de acordo com um determinado valor de tolerância, e identificar qual o sensor do grupo está defeituoso.

Os valores das métricas E_a e E_p indicam numericamente a proximidade das predições do sistema em relação aos valores corretos dos sensores, que também pode ser notada pelos gráficos comparativos das figuras. As métricas de sensibilidade indicam a possibilidade de degradação das predições na ocorrência de desvios em algum dos sensores. Sistemas com grande sensibilidade a desvios podem gerar predições que seguem os próprios desvios, o que impossibilita a detecção dos mesmos, ou que se tornam gradativamente mais distantes das leituras do sensor correspondente mesmo quando este se encontra em bom estado de funcionamento, acarretando em falsos alarmes de desvios. A tolerância aceitável para a sensibilidade dos sistemas pode ser definida de acordo com os resultados do algoritmo SPRT. O algoritmo SPRT deve gerar alarmes para níveis intoleráveis de desvios (definidos pelo valor de M) para o sensor defeituoso e gerar o mínimo possível de falsos alarmes para os demais sensores, de forma que não comprometa a identificação do sensor defeituoso.

Com exceção dos algoritmos do modelo SVM, utilizados por uma interface da biblioteca LIBSVM (CHANG; LIN, 2011), todos os demais algoritmos dos sistemas foram implementados em Matlab. Como o intuito dos ensaios é apenas verificar a aplicabilidade dos sistemas, não se implementaram as rotinas segundo as interfaces estabelecidas pela arquitetura OSA-CBM. Além disso, como dito na Seção 2.2, nenhuma rotina ou módulo de diagnóstico foi implementado. Neste trabalho tal tarefa é realizada a partir da interpretação dos resultados da detecção de mudanças das propriedades estatísticas dos resíduos.

Almejando verificar e entender o funcionamento dos sistemas, o primeiro ensaio foi realizado com um conjunto de dados gerados pela simulação do modelo de um poço. Aos corresponder as expectativas de aplicabilidade no primeiro ensaio, o segundo ensaio foi realizado com um conjunto de dados coletados de sensores reais de um poço de produção de petróleo.

4.3.1 Ensaios com Dados de Simulação

Os dados simulados foram gerados por um modelo de poço construído com simulador OLGA¹. O modelo representa um poço operado por *gas-lift* com um tubo de produção de 2500 metros de profundidade. A pressão do separador foi mantida em 1000 kPa e o reservatório foi ajustado com uma pressão de 18000 kPa. Os dados simulados foram coletados durante 100 horas de produção de poço, enquanto a vazão de gás injetado foi ajustado em

¹Mais informações sobre o simulador são encontradas em <http://www.sptgroup.com/Products/olga/>

diferentes valores, representando diferentes condições de operação. O conjunto de dados é composto pela pressão de fundo de poço (PT_f), no topo do tubo de produção (PT_t), no topo do anular (PT_a) e a montante do *choke* de injeção (PT_m), como ilustrado na Fig. 10. A taxa de amostragem é de uma amostra por minuto. Um ruído branco de desvio padrão equivalente a 30% do desvio padrão do sinal ($\sigma_N = 0.3\sigma_S$, onde σ_N é o desvio padrão do ruído, σ_S é o desvio padrão dos dados da variável correspondente e 0.3 é o valor próximo ao valor estimado do desvio padrão dos ruídos presentes nos dados reais da Seção 4.3.2) foi adicionado aos dados de cada sensor simulado. Os dados de cada uma das variáveis são apresentados na Fig. 11.

As 60 primeiras horas de dados foram usados para treinamento dos modelos AAKR e SVM. O restante dos dados foi dividido em duas partes: 20 horas como conjunto de otimização e 20 horas como conjunto de testes.

Para os Sistemas 1 e 3 foram utilizados o mesmo modelo AAKR, construído a partir dos dados de treinamento e cujos parâmetros são $h = 0.1$ e $n_m = 600$ (aproximadamente 17% das amostras do conjunto de treinamento). Os parâmetros escolhidos advêm dos resultados apresentados no gráfico da Fig. 12, onde os melhores parâmetros são aqueles que apresentam menor

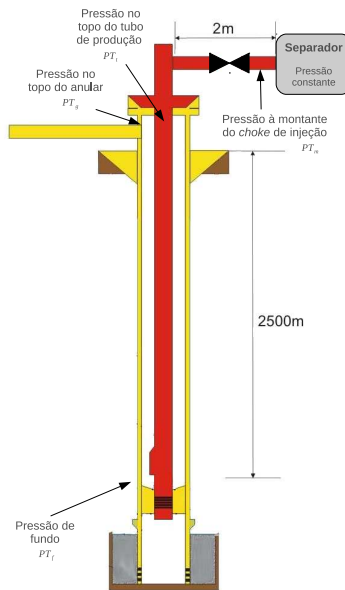


Figura 10 – Esquema do poço usado para gerar os dados de simulação.

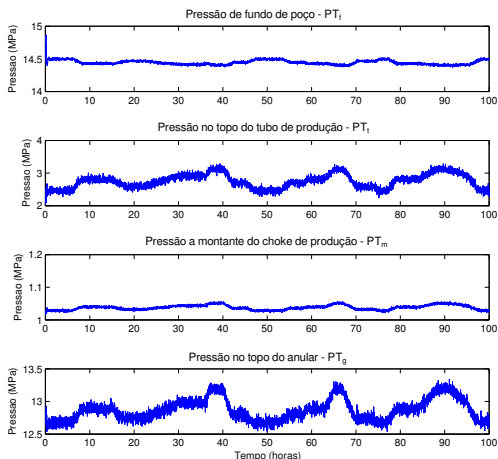


Figura 11 – Dados de variáveis gerados pela simulação de um poço de petróleo.

valor de MSE médio (ou E_a médio) entre o conjunto de dados de otimização e suas previsões gerados pelo modelo. Apesar de maiores quantidades de amostras resultarem em menores valores de MSE médio, a partir de 600 amostras o ganho de desempenho foi insignificante.

As configurações dos modelos SVM (um para cada variável monitorada) também foram obtidas a partir do treinamento com o conjunto de dados de treinamento e da minimização do MSE entre o conjunto de dados de otimização e suas previsões. Na Tabela 1 são apresentados as configurações dos modelos.

Para o Sistema 3, foi necessário determinar apenas os valores de σ_v e σ_q para cada uma das variáveis do modelo AAKR, uma vez que o modelo é o mesmo usado pelo Sistema 1. Inserindo desvios no conjunto de dados de

Tabela 1 – Configuração dos modelos SVM construídos com conjunto de dados de treinamento gerados por simulação.

Variável	ε	C	γ	Vetores suporte
PT_f	0.05	0.5	0.25	234
PT_t	0.1	1	8	246
PT_m	0.1	8	4	331
PT_g	0.12	32	2	316

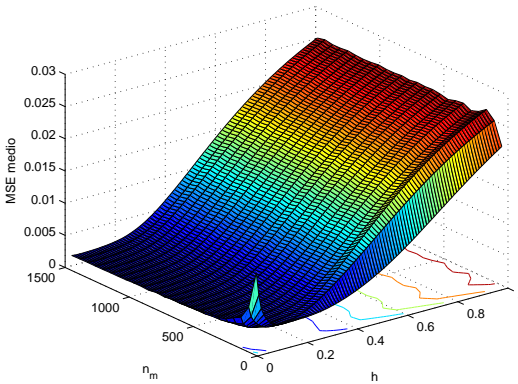


Figura 12 – Relação entre o MSE médio e diferentes valores de largura de banda h e vetores de memória n_m para o modelo AAKR na predição do conjunto de dados de otimização gerados por simulação.

otimização, foram selecionados os valores de σ_v e σ_q que geraram estimativas mais próximas dos desvios inseridos: $\sigma_v = 0.0001$ e $\sigma_q = 1$.

Na Tabela 2 são apresentados os resultados da análise de sensibilidade e acurácia dos sistemas, avaliados com o conjunto de dados de teste. Para tal análise, um desvio linearmente crescente terminando com uma magnitude de 3σ (três vezes o desvio padrão dos dados da variável) foi introduzido sequencialmente (em uma variável por vez) no conjunto de dados de teste.

Nota-se pela Tab. 2 que o Sistema 1 apresentou maior acurácia dentre os três sistemas, mas foi mais sensível a desvios quando comparado ao Sistema 3. Apesar de apresentar menor acurácia e maior sensibilidade cruzada dentre os sistemas, o Sistema 2 apresenta a vantagem de apresentar autossensibilidade nula, ou seja, as predições relativas a um sensor com desvio não

Tabela 2 – Valores médios das métricas de desempenho para os sistemas de monitoramento e validação de sensores: (E_a), autossensibilidade (S_A), sensibilidade cruzada (S_C) e erro de predição (E_p).

	E_a	S_A	S_C	E_p
Sistema 1	0.0020	0.2675	0.2682	0.0105
Sistema 2	0.0039		0.3356	0.0139
Sistema 3	0.0023	0.2002	0.1997	0.0046

sofrem influência desta falha. Apesar do melhor desempenho na presença de desvios, no Sistema 3 o filtro de Kalman acaba estimando pequenos desvios mesmo na ausência de desvios reais, o que implicou em menor acurácia em relação ao Sistema 1.

Na Figura 13 são ilustradas as predições dos três sistemas na condição de ocorrência de desvios nas leituras da variável PT_g . Nesta figura é possível notar as diferenças de desempenho retratadas pelos valores de autossensibilidade (S_A), sensibilidade cruzada (S_C) e erro de predição (E_p). O Sistema 1, com valores intermediários de S_C e E_p , aparentemente apresenta predições com menores E_p que o Sistema 2 para as variáveis PT_t e PT_m , mas piores que o Sistema 3; no caso da variável PT_f , o modelo regressivo do Sistema 2 parece ser mais independente dos valores de PT_g , uma vez que suas predições praticamente não foram afetadas pelos desvios presentes em PT_g . Com exceção das variáveis PT_f e PT_g , o Sistema 2 foi o mais afetado pelos desvios, confirmando seu maior valor de S_C , entretanto, as predições de PT_g permaneceram inalteradas pelos desvios, uma vez que a autossensibilidade do seu modelo regressivo é nula. O Sistema 3 apresenta, em geral, o melhor desempenho aparente entre os três sistemas, confirmando as métricas de desempenho, gerando predições muito próximas do desejado mesmo na ocorrência de desvios; entretanto, pelo valor de E_a , na ausência de desvios nas leituras o Sistema 3 tende a ser menos preciso que o Sistema 1.

Para a detecção automática de desvios e falhas, os algoritmos SPRT de cada um dos sistemas foram configurados a partir do conjunto de dados de otimização. Os valores de M foram definidos como os menores valores para os quais o algoritmo SPRT não gera nenhum falso alarme para os resíduos gerados por seus respectivos sistemas. A Tabela 3 contém os valores de M utilizados por cada um sistema e para cada uma das variáveis monitoradas.

Para o caso de ocorrência de desvios nas leituras da variável PT_g , por exemplo, os resultados da detecção automática pelos Sistemas 1 e 2 são ilustrados nas Figs. 14 e 15. Apesar dos falsos alarmes para as variáveis sem falhas, os Sistemas 1 e 2 detectaram corretamente a ocorrência de desvios

Tabela 3 – Configurações do algoritmo SPRT para o conjunto de dados de teste gerados por simulação.

	Valores de M por variável			
	PT_f	PT_t	PT_m	PT_g
Sistema 1	0.3	0.4	0.4	0.5
Sistema 2	0.2	0.45	0.5	0.6
Sistema 3	0.2	0.2	0.2	0.2

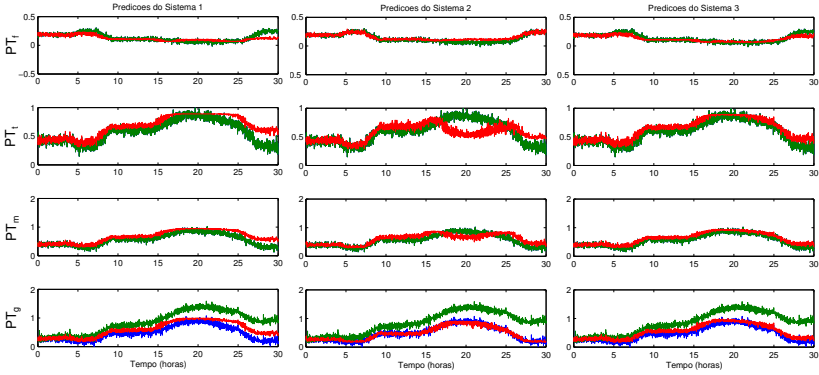


Figura 13 – Predições dos sistemas para o conjunto de dados de testes gerados por simulação, com inserção de desvios na variável PT_g . Linhas vermelhas são as predições, linhas verdes são as leituras das variáveis (sujeitas a falhas) usadas como entradas dos sistemas e linhas azuis são os valores verdadeiros das variáveis (diferentes das linhas verdes apenas na ocorrência de falhas). Todos os dados foram escalados no intervalo $[0, 1]$.

nas leituras de PT_g . O impacto dos falsos alarmes pode ser reduzido ou até mesmo eliminado usando outras regras simples sobre os resultados do algoritmo SPRT, como um limite para o número de decisões positivas (ocorrência de desvios) consecutivas. Por exemplo, poderia-se considerar a ocorrência de desvios apenas quando o algoritmo SPRT gerasse n ou mais resultados positivos consecutivos. Outra alternativa é a sintonia mais precisa dos parâmetros do algoritmo SPRT.

No caso do Sistema 3, a estimativa dos desvios sofreu menos interferência dos ruídos de leitura, como pode ser notado na Fig. 16. As correções efetuadas pelo filtro de Kalman contribuíram para a melhoria da qualidade das predições do modelo AAKR, gerando diferenças nítidas nas estimativas dos desvios das leituras de cada variável. Desta forma, o algoritmo SPRT conseguiu detectar corretamente e mais rapidamente a ocorrência de desvios, além de ter gerado menos falsos alarmes.

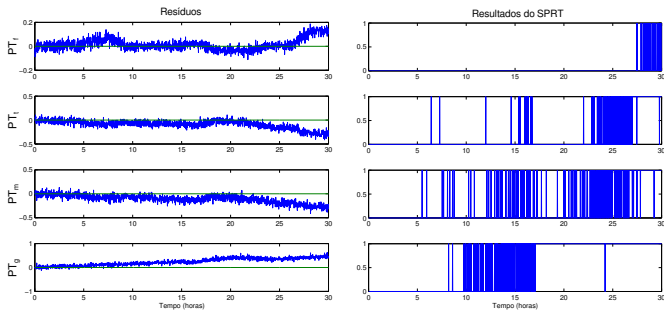


Figura 14 – Resíduos gerados pelo Sistema 1 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados de simulação.

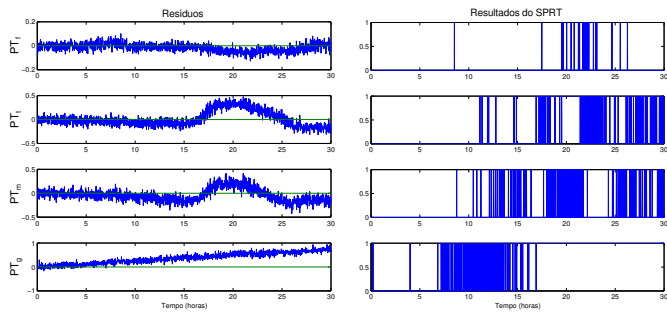


Figura 15 – Resíduos gerados pelo Sistema 2 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados de simulação.

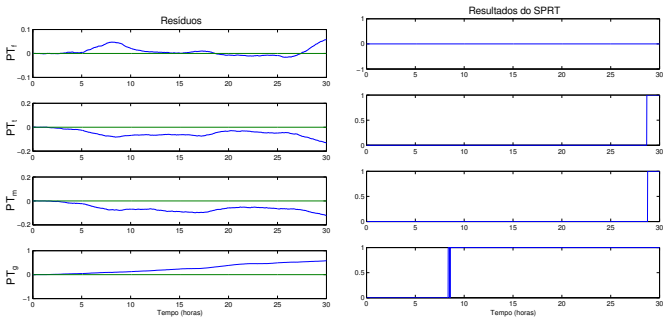


Figura 16 – Resíduos gerados pelo Sistema 3 e resultados do algoritmo SPRT para o caso de ocorrência de desvios em PT_g do conjunto de dados de simulação.

4.3.2 Ensaios com Dados Reais

O conjunto de dados reais é composto pelas leituras de pressão de três diferentes sensores de um poço de petróleo: um de fundo de poço (PDG), um na árvore de natal (TPT) e outro a montante do *choke* de injeção (PT_m). Os dados foram coletados durante 955 horas de operação de poço, amostrados a uma taxa de 1 amostra/minuto. Na Figura 17 são apresentados os gráficos de todo o conjunto de dados reais. Os visíveis *outliers* negativos representam, na verdade, leituras perdidas, não armazenadas por algum motivo desconhecido.

As primeiras 400 horas de dados foram separadas como conjunto de treinamento dos modelos, enquanto as 100 horas consecutivas foram usadas como conjunto de otimização e o restante como conjunto de testes. Por falta de conhecimentos mais profundos sobre os dados, nenhum procedimento de limpeza dos dados foi realizado.

Novamente, os Sistemas 1 e 2 compartilharam o mesmo modelo AAKR, construído com $h = 0.05$ e $n_m = 4800$ (20% do conjunto de dados de treinamento). Assim como feito com os dados de simulação, os parâmetros do modelo foram selecionados visando minimizar o MSE das predições do conjunto de dados de otimização. Neste caso, devido ao volume do conjunto de dados de treinamento, a quantidade n_m de amostras impactou significativamente no custo de tempo das predições. Considerou-se 4800 amostras uma quantidade razoável para a relação entre desempenho (valores de MSE)

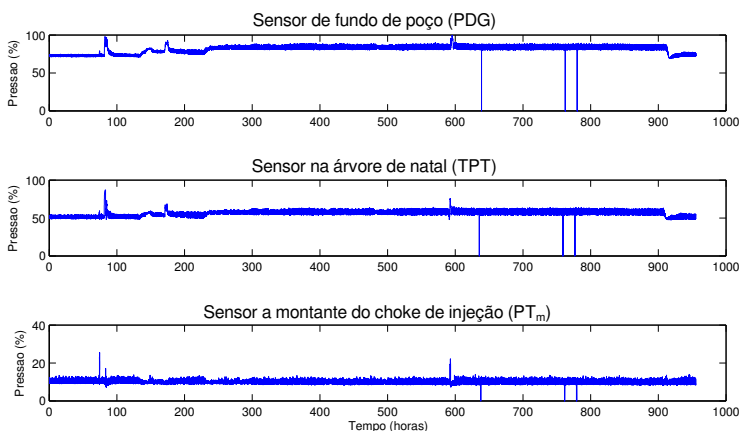


Figura 17 – Dados coletados por sensores de um poço de petróleo real, amostrados à taxa de uma amostra por minuto.

e custo das predições.

As configurações dos modelos SVM, obtidos a partir do conjunto de treinamento, são apresentadas na Tab. 4. Os valores dos parâmetros foram obtidos minimizando o MSE das estimativas do conjunto de otimização.

Para o Sistema 3, utilizando o mesmo modelo AAKR do Sistema 1, foram selecionados os parâmetros como $\sigma_v = 0.0001$ e $\sigma_q = 500$, cujos valores, encontrados por tentativa e erro, geraram as melhores estimativas de desvios inseridos no conjunto de dados de otimização. Em relação ao ensaio com dados de simulação, o valor mais alto de σ_q se deve aos sinais mais ruidosos dos dados reais, tornando as observações z dos desvios (vide Seção 3.3.1) menos confiáveis.

Na Tabela 5 são apresentados os resultados da análise de sensibilidade e acurácia dos sistemas, avaliados com o conjunto de dados de teste. Novamente, um desvio linearmente crescente terminando com uma magnitude de 3σ (três vezes o desvio padrão dos dados da variável) foi introduzido sequencialmente (em uma variável por vez) no conjunto de dados de teste.

Os resultados apresentados na Tab. 5 corroboram as análises dos sistemas com os dados gerados por simulação. Apesar de possuir autossensibilidade nula, ou seja, a insensibilidade do sistema para gerar estimativas de uma variável cujo sensor apresenta desvios, o Sistema 2 apresentou maior sensibilidade aos desvios quando estes estão presentes em alguma das entradas do modelo regressivo, como se pode constatar pelos valores mais altos de

Tabela 4 – Configuração dos modelos SVM construídos com conjunto de dados reais de treinamento.

Variável	ϵ	C	γ	Vetores suporte
PDG	0.12	256	512	5807
TPT	0.08	32	128	2286
PT_m	0.05	4	0.0625	4176

Tabela 5 – Valores médios das métricas de desempenho para os sistemas de monitoramento e validação de sensores: (E_a), autossensibilidade (S_A), sensibilidade cruzada (S_C) e erro de predição (E_p).

	E_a	S_A	S_C	E_p
Sistema 1	0.0035	0.3106	0.2806	0.0078
Sistema 2	0.0069		0.4754	0.0190
Sistema 3	0.0040	0.2418	0.2189	0.0046

E_a , E_p e S_C . Além disso, o Sistema 2 possui a menor acurácia, ou seja, gerou estimativas piores mesmo na ausência de falhas. O Sistema 1 apresentou as melhores estimativas na ausência de desvios, entretanto, apresentou maior sensibilidade em comparação ao Sistema 3. Apesar das correções realizadas pelo filtro de Kalman na presença de desvios, o Sistema 3 estima pequenos desvios a partir das variações dos resíduos mesmo na ausência de desvios reais nas leituras das variáveis, o que implicou em menor acurácia em relação ao Sistema 1.

As estimativas dos três sistemas na condição de ocorrência de desvios no sensor PDG são ilustradas na Fig. 18. Os Sistemas 1 e 3 apresentaram estimativas bem semelhantes, porém as correções efetuadas pelo filtro de Kalman no Sistema 3 tornaram as estimativas mais próximas das leituras originais, no caso dos sensores PDG e TPT, e mais estáveis, no caso do sensor PT_m . As estimativas do Sistema 2 apresentaram grande variância, mesmo no caso da estimação das leituras do sensor PDG, onde as entradas do modelo regressivo não apresentavam desvios. Além do próprio ruído original dos sensores, possivelmente as informações apenas de dois dos sensores disponíveis são insuficientes para inferir com melhor qualidade as leituras do terceiro.

Para a detecção automática de desvios e falhas, os algoritmos SPRT de cada um dos sistemas foram configurados a partir do conjunto de dados de

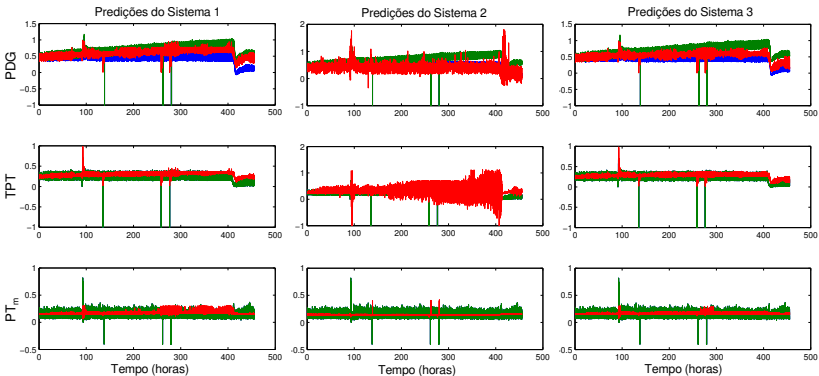


Figura 18 – Predições dos sistemas para o conjunto de dados reais de testes, com inserção de desvios nas leituras do PDG. Linhas vermelhas são as predições, linhas verdes são as leituras dos sensores (sujeitas a falhas) usadas como entradas dos sistemas e linhas azuis são as leituras originais dos sensores (diferentes das linhas verdes apenas na ocorrência de falhas). Todos os dados foram escalados no intervalo $[0, 1]$.

otimização. Os valores de M foram definidos como os menores valores para os quais o algoritmo SPRT não gera nenhum falso alarme para os resíduos gerados por seus respectivos sistemas. A Tabela 6 contém os valores de M utilizados por cada um sistema e para cada uma das variáveis monitoradas.

Para o caso de ocorrência de desvios no PDG, por exemplo, os resultados da detecção automática pelos Sistemas 1, 2 e 3 são ilustrados nas Figs. 19, 20 e 21, respectivamente. Pelos resultados do algoritmo SPRT na Fig. 19, o Sistema 1 detectou corretamente a ocorrência de desvios no PDG antes da ocorrência de vários falsos alarmes consecutivos nos demais sensores. Os falsos alarmes gerados próximos do instante 100 e 150 horas devem-se à uma possível anomalia (visível também próximo do instante 600 horas na Fig. 17) e a um *outlier*, citado anteriormente. Na verdade, estes resultados apontam a possibilidade de emprego do sistema para a indicação de outros tipos de falhas, como mudanças abruptas; entretanto, é de suma importância se certificar que o conjunto de treinamento dos modelos regressivos abrange o ponto de operação atual do poço.

Tabela 6 – Configurações do algoritmo SPRT para o conjunto de dados reais de teste.

	Valores de M por variável		
	PDG	TPT	PT _m
Sistema 1	0.4	0.4	0.4
Sistema 2	1.0	0.5	0.5
Sistema 3	0.1	0.1	0.1

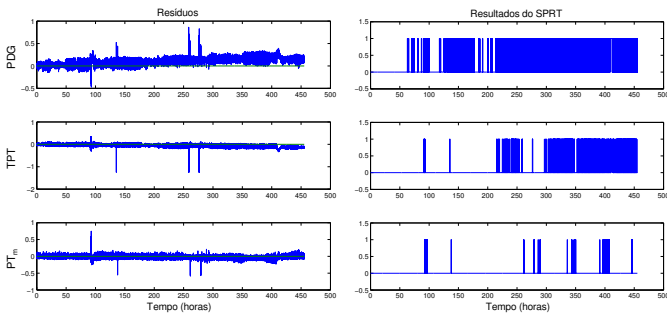


Figura 19 – Resíduos gerados pelo Sistema 1 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.

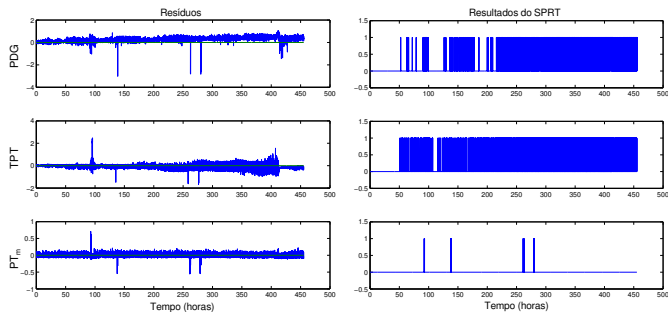


Figura 20 – Resíduos gerados pelo Sistema 2 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.

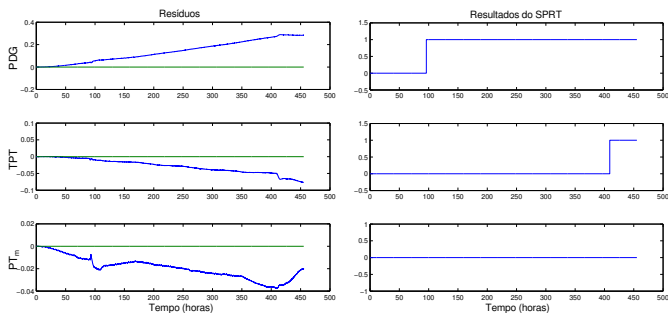


Figura 21 – Resíduos gerados pelo Sistema 3 e resultados do algoritmo SPRT para o caso de ocorrência de desvios nas leituras do PDG do conjunto de dados reais.

Os resultados da detecção pelo Sistema 2, apresentados na Fig. 20, refletem o pior desempenho do sistema verificado pelas análises de sensibilidade da Tab. 5. Apesar de aparentar média próxima de zero, a grande variância das estimativas das leituras do sensor TPT causaram a indicação incorreta de que o sensor com desvios seria o TPT. Aparentemente, a ocorrência de desvios no PDG afetou de tal forma as estimativas das leituras do TPT que a variância dos resíduos foi bem superior à variância dos resíduos nas estimativas do conjunto de otimização, usado para a configuração do algoritmo SPRT.

Pela Figura 21, apesar das estimativas de desvios no TPT e PT_m , nota-se que o Sistema 3 estimou corretamente o desvio verdadeiro no PDG, dada

a diferença de amplitude entre os desvios estimados. As estimativas dos desvios do PDG não conseguiram corrigir perfeitamente as leituras do sensor, pois desvios crescentes, apesar de menor amplitude, também foram estimados no TPT e no PT_m . Entretanto, as correções efetuadas pelo filtro de Kalman do sistema foram suficientes para uma indicação clara da ocorrência de desvios apenas no sensor PDG. No caso do Sistema 3, o monitoramento e detecção de outros tipos de falhas deve ser realizado pela análise dos resíduos (que neste caso não apontam desvios, uma vez que foram subtraídos das entradas do modelo) entre as estimativas geradas pelo modelo regressivo e as leituras dos sensores.

5 CONCLUSÕES

Neste trabalho foram apresentados o desenvolvimento e a implementação de sistemas de monitoramento e validação de sensores de poços de petróleo, utilizando técnicas de aprendizado de máquina para a construção de modelos empíricos baseados em histórico de dados¹. O emprego de tais sistemas visa garantir a qualidade e veracidade das informações coletadas pelos sensores, normalmente essenciais para ações de controle, otimização da produção e tomadas de decisões estratégicas para o negócio, além de preservar a segurança das operações.

Pensando na complexidade de todo o processo de produção petrolífero e na possibilidade de integração com sistemas de manutenção mais abrangentes, os sistemas propostos foram estruturados de acordo com uma arquitetura aberta para sistemas de manutenção baseados na condição de funcionamento (CBM), a OSA-CBM. Para o desenvolvimento dos modelos empíricos baseados em histórico, duas diferentes técnicas de regressão foram analisadas: a regressão por kernel auto-associativa (AAKR) e a *Support Vector Machines* (SVM). Foi apresentado um filtro de Kalman para a estimação de desvios, visando tanto um ganho de desempenho das predições dos modelos empíricos quanto uma alternativa para a detecção automática de desvios nos sensores. A detecção de desvios nos sensores foi realizado pelo algoritmo SPRT, através da análise de mudanças nas propriedades estatísticas dos resíduos ou das estimativas de desvios realizadas pelo filtro de Kalman.

Para verificar a aplicabilidade do trabalho desenvolvido, foram realizados ensaios com três diferentes sistemas de validação: o Sistema 1, composto por um modelo AAKR; o Sistema 2, composto por um modelo SVM; e o Sistema 3, composto por um modelo AAKR (devido aos melhores resultados nos ensaios quando comparado ao modelo SVM) e um filtro de Kalman. A avaliação dos sistemas considerou quatro diferentes métricas: acurácia (E_a), autossensibilidade (S_A), sensibilidade cruzada (S_C) e os erros de predição na presença de desvios (E_p). Em relação à detecção automática de desvios, avaliou-se os sistemas pela influência de falsos positivos na decisão final do sistema. Dois diferentes conjuntos de dados foram usados para os ensaios: um conjunto de dados gerados pela simulação do modelo de um poço de petróleo e um conjunto de dados coletados de sensores de um poço real.

Em ambos os ensaios, com dados simulados e reais, o Sistema 3 apresentou menor sensibilidade a desvios e o menor erro de predição E_p quando comparado ao Sistema 1, porém apresentou menor acurácia (maior E_a). Ape-

¹Parte deste trabalho foi publicado no *workshop* em automação e controle em plataformas de petróleo e gás, IFAC (BOECHAT; MORENO; HARAMURA, 2012).

sar das diferenças de desempenho, ambos os sistemas foram capazes de detectar corretamente a ocorrência de desvios. O Sistema 2, embora possua autosensibilidade nula de forma inerente, apresentou valores piores de E_a , E_p e S_C para os dois conjuntos de dados. No caso dos dados de simulação, o pior desempenho não impediu que o sistema detectasse corretamente a ocorrência de desvios. Entretanto, para o conjunto de dados reais, o Sistema 2 não foi capaz de gerar previsões consistentes e detectar a ocorrência de desvios corretamente, gerando uma quantidade inaceitável de falsos alarmes e impossibilitando a identificação do sensor defeituoso.

Embora os resultados apontem o emprego de modelos AAKR e do filtro de Kalman como promissores, não se pode descartar o uso de modelos SVM para sistemas de monitoramento e validação de sensores. É preciso considerar a pequena quantidade de sensores disponíveis no conjunto de dados reais. Possivelmente, em grupos maiores de sensores correlacionados os modelos SVM podem apresentar desempenhos melhores.

Tanto os sistemas compostos por modelos AAKR quanto por modelos SVM poderiam gerar melhores previsões com o emprego de técnicas automáticas de agrupamento ótimo de sensores, que deve ser considerado para trabalhos futuros. Além disso, a aplicação de métodos automáticos para validação dos próprios modelos empíricos seria desejável, uma vez que as características de um poço de petróleo mudam ao longo de sua vida útil, assim como o comportamento da correlação entre as variáveis mensuradas pelos sensores. Finalmente, vislumbra-se a possibilidade de construir modelos SVM de séries temporais, onde a saída do modelo não dependa apenas das entradas atuais, mas também de entradas anteriores (CHEN; CHANG; LIN, 2004).

REFERÊNCIAS BIBLIOGRÁFICAS

AGGREY, G.; DAVIES, D. Tracking the state and diagnosing Down Hole Permanent Sensors in Intelligent Well Completions with Artificial Neural Network. *Offshore Europe*, Society of Petroleum Engineers, 2007.

AN, S. H.; HEO, G.; CHANG, S. H. Detection of process anomalies using an improved statistical learning framework. *Expert Systems with Applications*, Elsevier Ltd, v. 38, n. 3, p. 1356–1363, 2011.

BOECHAT, A. A.; MORENO, U. F.; HARAMURA, D. J. On-Line Calibration Monitoring System Based on Data-Driven Model for Oil Well Sensors. In: LARS, I. (Ed.). *Proceedings of the IFAC Workshop on Automatic Control in Offshore Oil and Gas Production, Trondheim, Norway*. [S.l.: s.n.], 2012. p. 269–274.

BOECHAT, A. A.; MORENO, U. F.; TEIXEIRA, A. F. PROPOSTA DE UMA ARQUITETURA DE MONITORAMENTO BASEADO NA CONDIÇÃO DE PARA APLICAÇÃO EM UNIDADES MARÍTIMAS DE PRODUÇÃO. In: *VI Congresso Rio Automação, 2011, Rio de Janeiro*. [S.l.: s.n.], 2011.

CAMPOS, J. Development in the application of ICT in condition monitoring and maintenance. *Computers in Industry*, v. 60, n. 1, p. 1–20, 2009.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, p. 27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

CHEN, B.-J.; CHANG, M.-W.; LIN, C.-J. Load Forecasting Using Support Vector Machines: A Study on EUNITE Competition 2001. *IEEE Transactions on Power Systems*, v. 19, n. 4, p. 1821–1830, 2004.

CHERKASSKY, V.; MA, Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks : the official journal of the International Neural Network Society*, v. 17, n. 1, p. 113–26, 2004.

CHERKASSKY, V.; MULIER, F. *Learning from data: concepts, theory, and methods*. [S.l.]: John Wiley & Sons, 2007. (Wiley series on adaptive and learning systems for signal processing, communications, and control).

ECK, J. et al. Downhole monitoring: the story so far. *Oilfield Review*, v. 3, n. 3, p. 18–29, 1999.

GARVEY, J. et al. Validation of on-line monitoring techniques to nuclear plant data. *Nuclear Engineering and Technology*, Korean Nuclear Society, v. 39, n. 2, p. 133, 2007.

GRIBOK, A.; HINES, J. W.; UHRIG, R. Use of kernel based techniques for sensor validation in nuclear power plants. In: *American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Controls, and Human-Machine Interface Technologies (NPIC&HMIT 2000)*. Washington, DC: [s.n.], 2000.

HINES, J. W.; GARVEY, D. Development and application of fault detectability performance metrics for instrument calibration verification and anomaly detection. *Journal of Pattern Recognition Research*, v. 1, n. 1, p. 2–15, 2008.

HINES, J. W. et al. *Technical Review of On-Line Monitoring Techniques for Performance Assessment Vol. 2: Theoretical Issues*. U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research, 2008. v. 2.

HINES, J. W. et al. *Technical Review of On-Line Monitoring Techniques for Performance Assessment. Volume 3. Limiting Case Studies*. U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research, 2008.

HINES, J. W.; SEIBERT, R. *Technical Review of On-Line Monitoring Techniques for Performance Assessment. Volume 1. State-of-the-Art*. U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research, 2006. v. 1.

HINES, J. W.; UHRIG, R. Use of autoassociative neural networks for signal validation. *Journal of Intelligent and Robotic*, 1998.

HSU, C. et al. *A practical guide to support vector classification*. Department of Computer Science and Information Engineering, National Taiwan University, 2003.

MA, J.; JIANG, J. Applications of fault detection and diagnosis methods in nuclear power plants: A review. *Progress in Nuclear Energy*, Elsevier Ltd, v. 53, n. 3, p. 255–266, 2011.

Mauro Vitor de Oliveira. *Metodologia para validação de sinal usando modelos empíricos com técnicas de inteligência artificial aplicada a um reator nuclear*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, COPPE, 2005.

Penn State University; The Boeing Company. *Open Systems Architecture for Condition-based Maintenance (OSA-CBM) Primer*. Machinery Information Management Open Standards Alliance (MIMOSA), 2006.

SMOLA, A.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, n. 3, p. 199–222, 2004.

SOUZA, R. *Metodologia e desenvolvimento de um sistema de manutenção preditiva visando à melhoria da confiabilidade de ativos de usinas hidrelétricas*. Tese (Doutorado) — Universidade de Brasília, Brasília, 2008.

SWEARINGEN, K. et al. An Open System Architecture for Condition Based Maintenance Overview. In: *Aerospace Conference, 2007 IEEE*. [S.l.]: IEEE, 2007.

TAKRURI, M.; CHALLA, S.; YUNIS, R. Data fusion techniques for auto calibration in wireless sensor networks. In: *Information Fusion, 2009. FUSION'09. 12th International Conference on*. [S.l.]: IEEE, 2009. p. 132–139.

TAKRURI, M.; RAJASEGARAR, S.; CHALLA, S. Online drift correction in wireless sensor networks using spatio-temporal modeling. *Information Fusion, 2008 11th International Conference on*, IEEE, p. 1–8, 2008.

WALD, A. *Sequential analysis*. Wiley, 1947.

WAND, M.; JONES, M. *Kernel smoothing*. [S.l.]: Chapman & Hall/CRC, 1995.

ZAVAJJEVSKI, N.; GROSS, K. Sensor fault detection in nuclear power plants using multivariate state estimation technique and support vector machines. In: *Third International Conference of the Yugoslav Nuclear Society*. [S.l.: s.n.], 2000. p. 1–8.