LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

Na industria 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 em (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 em condição (CBM — *Condition Based Maintenance*) tendem à manutenção ótima, pois o desempenho dos instrumentos é monitorado durante a operação da planta e recalibrações físicas são realizadas apenas esse desempenho está degradado. 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). Porém, a redundância por *hardware* inclui a possível necessidade de sensores extras, além de apresentar dificuldades na detecção de sensores descalibrados que apresentam *drift* na mesma direção. Além disso, no caso dos sensores de fundo de poço, componentes redundantes ocupam um valioso e limitado espaço, 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 podem ser aplicados a quaisquer tipos de sistemas, sem a necessidade de conhecimentos específicos 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 diversas outros setores industriais.

1.1 DEFINIÇÃO DO PROBLEMA E ABORDAGENS EXISTENTES

TODO: Explicitar diferença entre predição e estimação.

Como apresentado em (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, continuo 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 afim 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 de desempenho de sensores baseados em modelos empíricos segue o esquema apresentado na Fig. 1. Supondo uma matriz de dados de treinamento $\mathbf{X} \in \mathbb{R}^{n \times p}$, onde n é 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 $\mathbf{r} \in \Re^{1 \times p}$ 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 saúde dos sensores (MA; JIANG, 2011).

Inicialmente vários trabalhos empregaram redes neurais para o desenvolvimento de modelos, mas atualmente as técnicas baseadas em kernel estão entre as mais utilizadas.

Em (HINES; UHRIG, 1998), um sistema de monitoramento de 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 de plantas nucleares, envolvendo problemas como *drifts* e erros grosseiros nos sensores.

Em (GRIBOK; HINES; UHRIG, 2000) foi realizada uma compara-

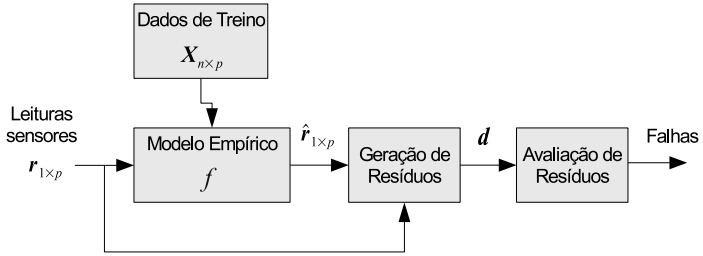

Figura 1 – Esquema de funcionamento de sistemas baseados em modelos empíricos para monitoramento de desempenho de sensores. Adaptado de (MA; JIANG, 2011).

ção 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.

Em (ZAVALJEVSKI; GROSS, 2000) foi realizada 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.

Em (GARVEY et al., 2007) foram apresentados os resultados da aplicação da técnica de regressão kernel auto-associativa (AAKR) a conjuntos de sensores de plantas nucleares. Diferentes métricas são apresentadas para a avaliação de desempenho dos modelos AAKR e a detecção de *drifts* é realizada pela análise de incerteza dos modelos.

Em (AGGREY; DAVIES, 2007) foi feito 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. A detecção de sensores degradados foi realizada por análises de limites (*thresholds*) dos valores dos resíduos.

Em (TAKRURI; RAJASEGARAR; CHALLA, 2008) foi proposto 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 da leitura correta, ou verdadeira, do sensor em questão.

Neste trabalho, os modelos empíricos foram desenvolvidos utilizando

AAKR ou SVM. A técnica AAKR possui implementação simples e esforço mínimo para o desenvolvimento de modelos, quando comparado às demais técnicas. A técnica SVM possui característica inferencial, o treinamento dos modelos é mais simples, comparado às redes neurais, e gera resultados estáveis.

1.2 OBJETIVOS E CONTRIBUIÇÕES

Esta dissertação tem como principais objetivos:

- Desenvolver um sistema de validação de 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 de validação de sensores para dados de poços de produção de petróleo;
- Apresentar um esquema de funcionamento do sistema de validação de sensores dentro de uma arquitetura mais genérica para sistemas de monitoramento e manutenção baseado em condição.

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 em 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

Conceitos gerais sobre monitoramento baseado em condição, importância de uma estrutura modular e bem definida.

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 desempneho de sistemas e o gerênciamento 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 materias 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 tornaria-se 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.

2.0.1 OSA-CBM

Existência da OSA-CBM e breve descrição sobre a respectiva arquitetura. Como o trabalho da dissertação, um sistema de monitoramento de calibração, se relaciona com a estrutura 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 dividido 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 6, 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. ?? é apresentado um esquema da organização dos módulos da arquitetura.

Figura do report da OSA 3.1.0 Primer

Os primeiros três blocos são específicos para determinadas tecnologias

e provêem as seguintes funções (Penn State University; The Boeing Company, 2006):

Copiar do artigo citado

A Fig. ?? ilustra um exemplo mais prático da estrutura de um sistema CBM.

Figura do artigo Souza2008, no artigo do rio automacao

2.0.2 Detecção e Correção de Drift sob a OSA-CBM

A maneira como os algoritmos de detecção de drift se inserem dentro da arquitetura OSA-CBM.

3 PROCESSAMENTO DE SINAIS PARA DETECÇÃO E CORREÇÃO DE DRIFT

TODO: Dividir as etapas em pré-processamento, processamento e pós-processmaneto. Incluir o KF na etapa de pós-processamento.

Pode-se dividir o processo de validação de sensores em três etapas principais: pré-processamento, processamento, pós-processamento dos sinais e o monitoramento de condição. Cada uma dessas etapas são detalhadas 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 são necessárias 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 $r_p \in \Re$ obtida de um sensor p é dada pela Eq. (3.1), onde $r_{norm,p}$ é o valor normalizado de r_p e $r_{max,p}$ e $r_{min,p}$ são os valores máximos e mínimos dos dados de treinamento relativos ao sensor p.

$$r_{norm,p} = \frac{r_p - r_{min,p}}{r_{max,p} - r_{min,p}} \tag{3.1}$$

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 predições 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 predições apresentam alta variância. Entretanto, para algumas técnicas de regressão, variáveis altamente correlacionadas podem causar singularidade na matriz de dados entrada e, consequentemente, predições de baixa qualidade. 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. Hines (HINES; SEIBERT, 2006) apresenta que agrupamentos

ótimos normalmente contém menos de 30 sensores e que a adição de sinais irrelevantes acaba por aumentar a variância das predições 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 predições 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, intrinsicamente relacionados, devem ser considerados para a seleção das amostras para o treinamento dos modelos: quantidade e representatividade. Geralmente, a quantidade de variáveis de entrada do modelo dita uma quantidade mínima de amostras necessárias para se “capturar” a correlação entre elas. Entretanto, essas amostras precisam representar as condições de operação que se espera encontrar quando o modelo estiver funcional.

3.2 PROCESSAMENTO DOS SINAIS

O processamento dos sinais é a etapa de predição propriamnente 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 $[r_1, \cdots, r_n]$ 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 $[r_1, \cdots, r_p]$ é usado para predizer os valores de um conjunto de variáveis de saída $[y_1, \cdots, y_q]$.

Um modelo auto-associativo é normalmente treinado para emular as próprias variáveis de entrada, ou seja, um conjunto de variáveis $[r_1, \cdots, r_p]$ é usado para gerar predições $[\hat{r}_1, \cdots, \hat{r}_p]$, que correspondem aos valores "corrigidos" das variáveis de entrada de acordo com o conjunto de dados de treinamento usado para a contrução 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 predições 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, $X_{i,j}$ 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

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_m,1} & X_{n_m,2} & \cdots & X_{n_m,p} \end{bmatrix}.$$

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

$$\mathbf{r} = \begin{bmatrix} r_1 & r_2 & \cdots & r_p \end{bmatrix}.$$

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, dada pela Eq. (3.2), onde d_i é a distancia Euclidiana entre o vetor de memória $\mathbf{X}_i$ e a entrada $\mathbf{r}$.

$$d_i(\mathbf{X}_i, \mathbf{r}) = \sqrt{(X_{i,1} - r_1)^2 + (X_{i,2} - r_2)^2 + \cdots (X_{i,p} - r_p)^2} \tag{3.2}$$

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

$$\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n_m} \end{bmatrix}.$$

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 w_i é o i-ésimo peso relativo a i-ésima distância d_i e h é a largura de banda da função kernel K.

$$w_i = K_h(d_i) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{d_i^2}{h^2}\right), \tag{3.3}$$

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 $j = 1, 2, \cdots, p$, obtém-se a saída do modelo, a predição $\hat{\mathbf{r}}$. Alternativamente, definindo $a = \sum_{i=1}^{n_m} w_i$, o estimador AAKR pode ser apresentado pela Eq. (3.5).

$$\hat{r}_j = \frac{\sum_{i=1}^{n_m} w_i \cdot X_{i,j}}{\sum_{i=1}^{n_m} w_i} \tag{3.4}$$

$$\hat{\mathbf{r}} = \frac{\mathbf{w}^T \mathbf{X}}{a} \tag{3.5}$$

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, normalmente existe um número crítico para cada conjunto de dados: quantidades menores pioram o desempenho do modelo drasticamente, quantidades maiores melhoram discretamente o desempenho ao custo de aumentos drásticos no tempo computacional (HINES et al., 2008b). 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.

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.

$$n_b = \frac{n_m}{2p} \tag{3.6}$$

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.

$$N_i = \sqrt{X_{i,1}^2 + X_{i,2}^2 + \cdots + X_{i,p}^2} \tag{3.7}$$

Um outra forma de realizar a seleção dos vetores de memória é a 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, uma matriz de dados de entrada $\mathbf{r} \in \Re^p$, de p variáveis independentes, é usada para estimar uma função $f(\mathbf{r})$ que correlaciona as p variáveis de $\mathbf{r}$ com uma variável dependente $y \in \Re$ a partir de n 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 $w \in X$ (espaço de características) e $b \in \Re$, é o estimador da função de dependência entre $\mathbf{r}$ e y. A Equação (3.8) pode ser reescrita 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".

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi^*) \tag{3.8}$$

$$\text{Sujeito a} \left\{ \begin{array}{rl} y_i - \langle \mathbf{w}, \mathbf{r}_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{r}_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{array} \right\}$$

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

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 ξ_i e ξ_i^* são chamadas de variáveis de folga, que determinam o grau de penalização aplicados às estimativas com erros maiores que ε. A Figura 2 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 ξ_i, ξ_i^* na função objetiva, 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.

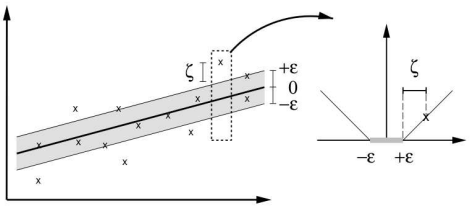

Figura 2 – Zona de insensibilidade e penalização no modelo SVM (SMOLA; SCHöLKOPF, 2004).

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).

$$W(\alpha^*, \alpha) = -\varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(r_i, r_j) \tag{3.10}$$

$$\text{Sujeito a} \left\{ \begin{array}{c} \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C \end{array} \right\}$$

A função $f(\mathbf{r})$ passa a ser dada pela Eq. (3.11), onde α, α^* são multiplicadores de Lagrange e $K(\mathbf{r}, \mathbf{r}_i)$ é a função kernel que substitui os produtos internos dos dados de entrada e permite lidar com não linearidades eventualmente presente nos dados. Os vetores $\mathbf{r}_i$ são chamados de "vetores suporte" e N_{sv} (normalmente $N_{sv} << n$) é o número de vetores suporte corresponden-

tes à quantidade de valores de y que estão pelo menos ε afastados de suas estimativas $f(\mathbf{r})$.

$$f(\mathbf{r}) = \sum_{i=1}^{N_{sv}} (\alpha_i^* - \alpha_i)\, K(\mathbf{r}, \mathbf{r}_i) + b \tag{3.11}$$

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 α_i^*, α_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. Uma função kernel largamente utilizada para regressão é a função de base radial (RBF), apresentada na Eq. (3.12), onde γ é a largura de banda, o parâmetro de configuração.

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

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, constroi-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, o problema de sobreajustamento (*overfitting*) dos dados poderá ocorrer, incluindo ruídos na modelagem. Considerando o uso de funções kernel do tipo Gaussiana ou RBF, a complexidade dos modelos SVM é determinada pelos parâmetros C, ε e γ; nos modelos AAKR, pelo parâmetro h. Em muitas aplicações, esses valores podem são encontrados por tentativa e erro, ou *grid search*, onde treinam-se vários modelos com diferentes permutações de valores de configuração e avalia-se a qualidade das predições destes atráves 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.

A construção de um modelo empírico normalmente busca minimizar o erro (ou risco de predição) entre a predição $f(\mathbf{r})$ do modelo e a saída y do sistema para uma dada entrada $\mathbf{r}$. A medida de qualidade que normalmente se adota é a minimização do erro quadrático médio (MSE), como a Eq. (3.13), onde n é a quantidade de observações utilizadas para a construção do modelo.

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - f(\mathbf{r}_i)) \tag{3.13}$$

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.

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.

Como dito na Seção 1.1, os resíduos correspondem à diferença entre

as predições do modelos e as medições dos sensores correspondentes. Considerando o sistema sensoriado 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, como apresentado na Seção 3.4.1.

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

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.

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

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 descorrelação 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.

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

Quando uma nova observação r e a estimação de seu valor corrigido $\hat{r}$ estiverem disponíveis, os seguintes passos são executados para se obter uma

nova estimação do desvio $d^{(k)}$ eventualmente presente no sensor:

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

2. MSE mínimo da predição,
$$M^{(k|k-1)} = M^{(k-1)} + \sigma_v$$

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

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

3.4.1 Avaliação dos Resídos e Detecção de Anomalias/Drift

TODO: SPRT, detecção pela incerteza. Present some possible methods to detect anomalies, like autocorrelation, densidade de potência etc.

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 $f(\mathbf{r})$ é a estimativa da medida y, medida produzida pelo sensor que se deseja monitorar, a partir das medidas $\mathbf{r}$ de sensores correlacionados.

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

Considerando o sistema sensoriado 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 pedições 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 passíveis de detecção, neste trabalho emprega-se o método SPRT, descrito a seguir.

3.4.2 SPRT

TODO: Seguir artigo IFAC

4 APLICAÇÃO DO SISTEMA IMPLEMENTADO A DADOS DE POÇOS DE PETRÓLEO

Neste capítulo são apresentados os sistemas implementados para monitoramento e validação de sensores de poços de petróleo, além dos resultados da avaliação destes sistemas para diferentes conjuntos de dados, gerados a partir de simulações ou a partir de sensores reais.

4.1 SISTEMAS DE VALIDAÇÃO DE SENSORES

TODO: 3 ou 4 sistemas serão implementados? Figuras e diagramas

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

4.1.1 Sistema 1 — AAKR-SPRT

4.1.2 sistema 2 — AAKR-KF-SPRT

4.1.3 Sistema 3 — SVM-SPRT

4.2 AVALIAÇÃO

Uma vez que o modelo tenha sido treinado e otimizado com o conjunto de dados de treinamento, o conjunto de validação é utilizado para avaliar o desempenho do modelo desenvolvido. Segundo Hines (HINES; GARVEY, 2008), tradicionalmente o desempenho de sistemas de monitoramento de calibração de sensores é mensurada a partir de três indicadores: acurácia, auto-sensibilidade e sensibilidade cruzada. A acurácia mensura a habilidade do modelo para gerar predições corretas e precisas das leituras dos sensores. A auto-sensibilidade indica a habilidade do modelo para gerar predições 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 predições dos demais sensores.

Apesar desses indicadores qualificarem características essenciais de todo sistema de monitoramento de calibração, a inspeção visual das predi-

ções é de fundamental importância. Modelos com bons índices de qualidade podem gerar predições insatisfatórias, como sinais muito suavizados, representando uma estimação muito grosseira da realidade e, por isso, sem valor prático para a tomada de decisões.

É importante destacar ainda a possível necessidade de retreinamento do modelo com dados mais atualizados da planta. Mudanças no ponto de operação, nas condições de equipamentos ou nas características do meio podem afetar significativamente na forma como as variáveis mensuradas pelos sensores se correlacionam. Assim, dependendo da instensidade da mudança, dados atualizados podem ser simplesmete incorporados ao conjunto de treinamento anterior ou devem compor um conjunto completamente novo.

4.3 DESCRIÇÃO DOS DADOS

Seção semelhante a do artigo submetido para o ifac.

4.3.0.1 Simulação

4.3.0.2 Dados Reais

4.4 ENSAIO — CONSTRUÇÃO DOS MODELOS EMPÍRICOS

Testes sem o filtro de Kalman com AAKR e SVM. Verificar qualidade das predições. Testes do AAKR com KF. Se possível, testar o SVM também, pelo menos para algumas variáveis. Testes com sinais filtrados (fácil para o caso de simulação).

4.4.1 Dados de Simulação

4.4.1.1 AAKR

O resultado do processo de otimização dos parâmetros do modelo AAKR é apresentado na Fig. 3. Modelos com valores altos de h apresentaram grandes valores de MSE total, independente da quantidade de vetores de memória, n_m. Maiores valores de n_m combinados com menores valores de h tenderam a apresentar modelos com menor MSE total. Porém, apesar de não ser perceptível pela Fig. 3, para valores muito pequenos de h ocorreram

leves acréscimos no MSE total, para todos as quantidades n_m. Para evitar *overfitting* e o armazenamento de grandes quantidades de dados desnecessariamente, escolheu-se $h = 0.1$ e $n_m = 600$ (20% da quantidade amostras dos dados de treinamento).

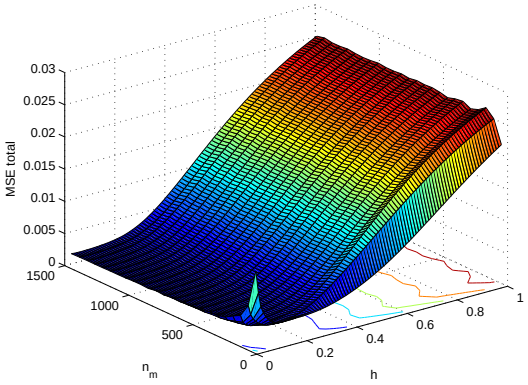

Figura 3 – Relação entre o MSE total e diferentes valores de largura de banda h e vetores de memória n_m para o modelo AAKR aplicado aos dados de simulação para otimização.

4.4.1.2 SVM

Os resultados da otimização dos parâmetros do modelo SVM são apresentados nas Tabelas 1, 2, 3 e 4. Para cada valor de ε, os valores de C e γ são obtidos por busca em grade, variando-se os valores em potências de 2. Nota-se que o valor de ε parece ser o principal determinante da quantidade de vetores suporte armazenados pelo modelo. Menores valores de ε implicaram em maiores números de vetores suporte e menores valores do MSE. Entretanto, maior número de vetores suporte indicam modelos mais complexos, e modelos excessivamente complexos tendem a incluir ruídos na modelagem. Portanto, modelos com a mínima complexidade possível para se atingir valores de MSE toleráveis são normalmente as melhores opções. Os modelos selecionados para o sistema de validação estão destacados nas Tabelas 1, 2, 3 e 4.

Tabela 1 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição da variável PT_f dos dados de simulação.

ε	C	γ	Vetores suporte	MSE ($\times 10^{-3}$)
0.03	128	4	122	0.19
0.04	512	8	73	0.26
0.06	0.125	0.5	33	0.42
0.08	1	1	25	0.66
0.1	256	2	20	0.93
0.5	0.125	16	0	33.19

Tabela 2 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição da variável PT_t dos dados de simulação.

ε	C	γ	Vetores suporte	MSE ($\times 10^{-3}$)
0.1	32	32	473	4.53
0.2	2048	0.5	22	4.74
0.22	1024	1	15	4.92
0.25	2048	0.25	9	5.43
0.3	16	0.25	3	7.34
0.5	0.25	4	3	32.63

Tabela 3 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição da variável PT_m dos dados de simulação.

ε	C	γ	Vetores suporte	MSE ($\times 10^{-5}$)
0.005	0.125	1024	280	1.068
0.008	0.125	128	38	1.057
0.009	0.125	32	15	1.054
0.01	0.125	0.125	4	1.095

4.4.2 Dados Reais

4.4.2.1 AAKR

O resultado do processo de otimização dos parâmetros do modelo AAKR é apresentado na Fig. 4. Para este conjunto de dados, menores valores

Tabela 4 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição da variável PT_g dos dados de simulação.

ε	C	γ	Vetores suporte	MSE ($\times 10^{-3}$)
0.1	8	32	220	3.23
0.15	512	0.5	31	3.31
0.2	512	0.125	9	4.05
0.5	0.125	8	0	28.17

de h realmente geraram menores valores de MSE total, exceto para quantidades n_m muito pequenas. Entretanto, modelos com $h < 0.07$ apresentaram problemas numéricos para a predição dos dados de otimização, pois valores muito pequenos de h geram modelos com baixa capacidade de generalização, onde os denominadores da Eq. ?? são muito próximos de zero. Para diminuir a possibilidade de problemas numéricos por baixa capacidade de generalização e ao mesmo tempo obter baixos valores de MSE total, adotou-se $h = 0.15$.

Em relação ao número de vetores de memória, foi escolhido $n_m = 4800$, que corresponde a 20% do conjunto de dados de treinamento. Maiores valores de n_m apresentaram impacto insignificante na redução do MSE (alterações a partir da quarta casa decimal).

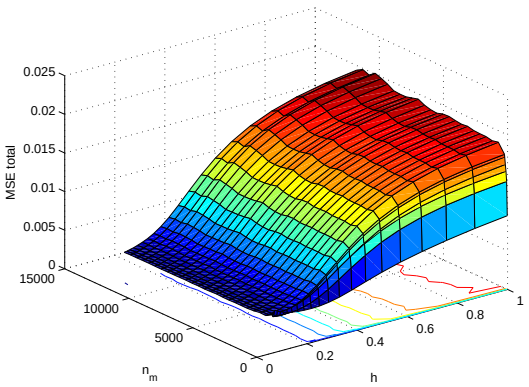

Figura 4 – Relação entre o MSE total e diferentes valores de largura de banda h e vetores de memória n_m para o modelo AAKR aplicado aos dados reais para otimização.

4.4.2.2 SVM

As Figuras ?? apresentam o desempenho, em termos do MSE, de diferentes modelos SVM, variando os parâmetros C, γ e ε.

Tabela 5 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição do sensor PDG dos dados reais.

ε	C	γ	Vetores suporte	MSE
0.5	0.5	1	19621	7.19
0.8	32	0.25	17109	7.12
1	0.25	0.5	15886	6.97
2	0.5	2	10662	7.00
5	0.25	2	2221	8.86

Tabela 6 – Relação entre parâmetros do modelo SVM e o impacto no MSE na predição do sensor TPT dos dados reais.

ε	C	γ	Vetores suporte	MSE
0.8	0.03125	0.03125	16775	4.36
2	0.0625	0.5	6922	4.25

4.4.2.3 Comentários

Para ambas as técnicas, o processo de otimização dos modelos norteada pelo MSE foi fortemente influenciada pela largura de banda da função kernel. Nos modelos AAKR, o MSE total cresceu com o aumento dos valores de h, entretanto, valores muito pequenos ocasionaram problemas numéricos. Nos modelos SVM,

Não houve uma escolha automática das variáveis, usou-se senso de engenharia.

4.5 ENSAIO — VERIFICAÇÃO DE ACURÁCIA

4.5.1 Dados de Simulação

4.5.1.1 AAKR

4.5.1.2 SVM

4.6 ENSAIO — ANÁLISE DE SENSIBILIDADE E DETECÇÃO DE *DRIFTS*

4.6.1 Dados Reais

4.6.1.1 AAKR

4.6.1.2 SVM

4.7 ENSAIO — ESTIMAÇÃO DE *DRIFTS* COM O KF

5 CONCLUSÕES

Revisão da importância do monitoramento de calibração. Revisão da sistema implementado e sua relação com a estrutura OSA-CBM. Revisão dos experimentos realizados e os resultados obtidos. Resumir as vantagens e desvantagens da abordagem escolhida, analisar de forma crítica as contribuições da dissertação.

Perspectivas de trabalhos futuros.

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.

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

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, jan. 2004. ISSN 0893-6080. <http://www.ncbi.nlm.nih.gov/pubmed/14690712>.

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).

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, November*. [S.l.: 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. <http://jprr.org/index.php/jprr/article/view/5>.

HINES, J. W. et al. *Technical Review of On-Line Monitoring Techniques for Performance Assessment Vol. 2: Theoretical Issues*. Washington, DC, 2008. v. 2.

HINES, J. W. et al. *Technical Review of On-Line Monitoring Techniques for Performance Assessment. Volume 3. Limiting Case Studies*. Washington, DC, 2008.

HINES, J. W.; SEIBERT, R. *Technical Review of On-Line Monitoring Techniques for Performance Assessment. Volume 1. State-of-the-Art*. Washington, DC, 2006. v. 1.

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

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*. [S.l.], 2006.

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

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.

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

ZAVALJEVSKI, 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.