CLASSIFICADORES HIERÁRQUICOS E PCA NA DETECÇÃO E IDENTIFICAÇÃO DE FALHAS EM REDES DE ESCOAMENTO

Carlos André Vaz Junior[1] (EQ/UFRJ), Ofélia de Q. F. Araújo (EQ/UFRJ), José Luiz de Medeiros (EQ/UFRJ)

[1]Escola de Química da UFRJ, Bloco E, CT, Ilha do Fundão, CEP: 21949-900 Rio de Janeiro, RJ,
cavazjunior@gmail.com

O campo de instrumentação voltada para indústria química evoluiu fortemente nas últimas décadas, com redução no custo de componentes de sistemas eletrônicos e aumento do número de aplicações nos quais podem ser empregados. Tal progresso vem promovendo disponibilidade crescente de dados na forma de séries temporais para a monitoração e controle do processo. Extrair informações relevantes a partir da massa total de dados torna-se um importante campo de estudo. Metodologias tais como "redes neurais", "análise dos componentes principais" (PCA), "k-means" e "classificação hierárquica" são técnicas de mineração de dados com vastas aplicações em tratamento de dados, análise de clusters e extração de padrões. Essas ferramentas permitem tratar a crescente disponibilidade de dados provenientes de plantas químicas e petroquímicas altamente instrumentadas. Neste trabalho, foram desenvolvidas duas metodologias com a finalidade de detectar e identificar falhas em rede de dutos para transporte de nafta. Enquanto a primeira baseia-se em classificadores hierárquicos, a outra aborda o problema por análise dos componentes principais. A classificação hierárquica atua em conjunto de dados multidimensionais, e tem por objetivo definir partições ótimas. Para tanto, se utiliza um critério pré-estabelecido para medir similaridade entre elementos. As classes das partições resultantes devem ser homogêneas e bem separadas, isto é, devem satisfazer a condição de que elementos de uma mesma classe sejam semelhantes entre si. Métricas baseadas em distância "Euclidiana", "City block" e "Minkowski" foram aqui aplicadas. A ferramenta de classificação mostrou-se capaz de resolver com elevado grau de acerto conjuntos de dados com duas classes: "operação normal" e "operação anormal". Os classificadores também foram empregados com sucesso na diferenciação entre "falhas de sensores" e "ocorrências de vazamentos". O êxito no processo mostrou-se fortemente influenciado pela severidade da anomalia investigada e pela presença de ruído inerente à base de dados utilizada. Por outro lado, a abordagem por PCA baseia-se na existência de elevada correlação entre as leituras dos sensores de pressão e vazão ao longo da rede, sendo tal correlação modificada diante da ocorrência de anormalidades. Em decorrência, modelos PCA ajustados para dados com e sem falhas apresentam distinções significativas entre si. Métricas de similaridade Spca("PCA similarity factor") e Sdist("PCA distance similarity factor"), assim como índices baseados no erro quadrático médio de previsão, foram aqui empregadas. As abordagens propostas, especialmente a métrica Sdist, mostraram-se eficientes na detecção de anomalias de variadas severidades. Algum grau de precisão também foi obtido para a localização do sensor em falha.

detecção de falhas, análise de componentes principais, classificação hierárquica, séries temporais.

1. INTRODUÇÃO

A disponibilidade crescente de observações temporais geradas a partir de sensores industriais atualmente em uso criou um novo desafio: extrair da massa total de dados informações relevantes acerca do processo. Neste trabalho foram desenvolvidas duas abordagens distintas e complementares com a finalidade de detectar anormalidades tais como vazamentos e falhas em sensores numa rede de dutos para transporte de nafta. A primeira metodologia é baseada em técnica de estatística multivariável denominada "análise dos componentes principais" (PCA). Propriedades físicas inerentes ao processo de escoamento promovem elevada correlação entre o comportamento dos dados de pressão e vazão gerados pelos sensores posicionados ao longo da rede. Tal correlação, porém, é modificada diante da ocorrência de uma anomalia, sendo este o princípio básico da detecção de falhas via PCA. Complementarmente à aproximação por PCA, uma segunda abordagem, baseada em classificadores hierárquicos, foi empregada. O classificador hierárquico busca separar diferentes observações temporais em agrupamentos, usando para isso critério de similaridade previamente definido.

Enquanto que a análise por componentes principais mostra-se eficaz na identificação de modificações no padrão de comportamento do sistema e, portanto, na detecção de anomalias, esta abordagem por si só não é capaz de indicar com precisão o momento exato de início da falha. Para que essa identificação possa ser feita com maior precisão, as observações temporais anteriores e posteriores a ocorrência da anomalia são submetidas a um classificador, sendo então alocadas em dois grandes agrupamentos, ou "clusters". Enquanto pontos sem anomalia situam-se em um agrupamento, amostras representantes de instantes com falha operacional encontram-se no outro. A técnica permite ainda estimar a severidade da falha a partir do afastamento espacial dos clusters

formados. A abordagem híbrida busca assim unir as vantagens provenientes das duas técnicas de reconhecimento de padrões bem estabelecidas.

2. METODOLOGIA

2.1 Rede de Escoamento

Com o objetivo de desenvolver e testar ferramentas de detecção e identificação de anomalias utilizou-se o simulador de redes de transporte de fluidos proposto por Vaz Junior (2006). A simulação computacional permite a realização de um vasto número de experimentos envolvendo a dutovia, possibilitando que diversos tipos de falhas sejam implementadas. A rede aqui simulada tem como objetivo o transporte de nafta ao longo de uma extensão de 30 quilômetros ligando o vértice 1 (origem) ao vértice 8 (destino). A Figura 1 esquematiza a rede e expõe a nomenclatura para tubos e vértices adotada

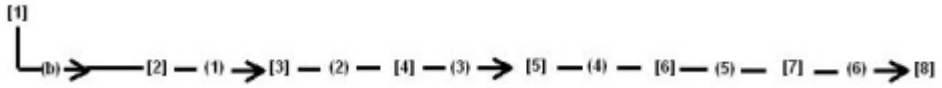


Figura 1. Rede de escoamento [x]: vértice, (b): bomba, (x): tubo

A pressurização e tráfego no interior do sistema são definidos via especificação de pressão (P) e de vazão mássica externa (W) diretamente nos vértices 1 e 8. Os demais parâmetros são calculados a partir de equações fenomenológicas de mecânica dos fluidos. Um sinal de ruído aleatório é incorporado aos dados gerados pelo simulador. As séries temporais obtidas possuem intervalo de 1 minuto entre cada observação. A Tabela 1 apresenta os sensores disponíveis, juntamente com sua localização e tipo.

Tabela 1. Localização e tipo dos sensores instalados onde [] representa vértice e () representa tubo

Sensor	Localização	Tipo	Sensor	Localização	Tipo
1	[1]	Vazão	9	[7]	Pressão
2	[8]	Vazão	10	[8]	Pressão
3	[1]	Pressão	11	(1)	Vazão
4	[2]	Pressão	12	(2)	Vazão
5	[3]	Pressão	13	(3)	Vazão
6	[4]	Pressão	14	(4)	Vazão
7	[5]	Pressão	15	(5)	Vazão
8	[6]	Pressão	16	(6)	Vazão

2.2 Metodologia de cálculo de PCA

De forma resumida, a metodologia da técnica PCA aplicada pode ser descrita como a obtenção dos autovetores e autovalores da matriz de covariância obtida a partir dos dados normalizados. Os autovetores, especialmente aqueles com maiores autovalores associados, provêem importantes informações sobre o padrão de distribuição dos dados (Hart e Duda, 2000). Desse modo, ordenando os autovalores é possível deduzir a quantidade de informação descrita por cada autovetor. Neste trabalho utilizou-se número de autovetores suficientes para representar 99% do comportamento dos dados. Denominando os autovetores de "componentes", aqueles associados aos autovalores mais elevados são então chamados "componentes principais". Os componentes principais compõem a matriz de componentes (MC). Na prática, tornar-se mais fácil representar o modelo PCA através da matriz "C", conforme Equação 1. A matriz pode então ser aplicada diretamente aos dados experimentais ($\hat{x}\exp_k$) obtendo-se os dados modelados ($\hat{x}\mathrm{mod}_k$), de acordo com a Equação 2

$$C=MC*MC^T \quad (1)$$

$$\hat{x}\mathrm{mod}_k = C\,\hat{x}\exp_k^T \quad (2)$$

2.3 Metodologia de detecção de anomalias via Q e SPE

Nesta abordagem, detecta-se falha quando o erro de previsão do modelo extrapola os limites estabelecidos pelos intervalos de confiança. O erro foi calculado através dos parâmetros SPE e Q. A métrica Q, apresentada

por Wise e Gallagher (1996), baseia-se no cálculo do erro quadrático de predição, ou seja, na diferença entre o valor predito pelo modelo e o observado. Enquanto a métrica SPE representa o erro quadrático da previsão ponderado com o tempo (square prediction error) (Dunia et al, 1996). Esta métrica possui "efeito memória", ou seja, erros passados tem maior ou menor influência sobre o valor mais recente. Tal ponderação é efetuada através do parâmetro l. Quanto mais próximo do valor unitário lambda encontra-se, menor a influência do passado, sendo SPE função principalmente dos valores mais recentes

Para a aplicação desta metodologia, as séries temporais geradas pelo simulador foram separadas em duas fases, sendo a primeira denominada "Fase A" ou "fase de treino". Nesta, ajustam-se os parâmetros do modelo PCA a série temporal sem falhas. Após o término desta, inicia-se a "Fase B" ou "fase de vigília". Nesta, implementa-se então a anomalia que deverá ser detectada. A Figura 2 representa a distribuição temporal das fases e do instante a partir do qual ocorre anomalia. Valores de Q e/ou SPE anormalmente elevados durante a fase B são indícios da ocorrência de falha.

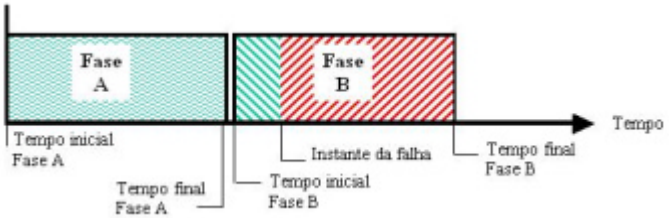


Figura 2. Representação esquemática do posicionamento temporal das fases A e B

2.4 Metodologia de detecção de anomalias via Spca e Sdist

Esta abordagem também se baseia na comparação das fases A e B, através da aplicação de índices de "simetria" ou "similaridade". Reduzida similaridade entre as duas fases indica a ocorrência de falha na fase de vigília. Duas métricas de similaridade foram utilizadas: Spca e Sdist. Para Spca, considera-se cada fase como um conjunto de dados. Assume-se que o modelo PCA que descreve cada conjunto é composto por "k" componentes principais. A similaridade entre os dois grupos de dados é quantificada através da comparação de seus componentes principais. Ou seja, através do cálculo do ângulo entre os autovetores (Equação 3), sendo é o ângulo formado entre o "i-ésimo" componente principal do primeiro conjunto de dados e o "j-ésimo" componente principal do segundo conjunto (Singhal e Seborg, 2002).

$$Spca = \frac{1}{k}\sum_{i=1}^{k}\ \sum_{j=1}^{k}\cos^2 q_{ij} \tag{3}$$

Enquanto o fator de similaridade Spca é influenciado pela orientação espacial do subespaço gerado pelos componentes principais, o fator Sdist (ou "distance similarity factor") é usado para situações onde os conjuntos de dados têm orientação espacial similar, mas estão localizados em posições distintas. O fator Sdist é a probabilidade que o centro do conjunto de dados A (f) esteja ao menos a distância do centro dos dados B ($\overline{x}_B$). Calcula-se Sdist através da Equação 4 (Singhal e Seborg, 2002)

$$Sdist = 2\left[1-\frac{1}{\sqrt{2p}}\int_{-\infty}^{f} e^{-\frac{z^2}{2}}dz\right], \text{onde}\quad f=\sqrt{(\overline{x}_A-\overline{x}_B)\sum\nolimits_{s}^{*-1}(\overline{x}_A-\overline{x}_B)^T} \tag{4}$$

A aplicação das métricas Spca e Sdist foi efetuada através de duas aproximações distintas, denominadas de "estática" e "dinâmica". Na abordagem estática, comparam-se os dados obtidos em duas janelas temporais distintas - "fase A" e "fase B" (Figura 2). Caso não exista falha na fase de vigília, o resultado esperado é que as métricas aproximem-se do valor unitário, indicando elevada similaridade. Já na abordagem dinâmica, a fase B torna-se móvel, ou seja, desloca-se ao longo do eixo temporal, afastando-se gradualmente do término da fase A (Figura 3). O afastamento da fase de vigília prossegue até a formação da última fase B, que começará no instante D e se estenderá até o instante E. A Figura 3 representa ainda a ocorrência de falha a partir do instante F. Observa-se que, a medida que a fase B desloca-se, a falha vai preenchendo uma parcela crescente desta. Esse fenômeno, não observado na abordagem estática, melhora o grau de detecção da ferramenta (Singhal e Seborg, 2002).

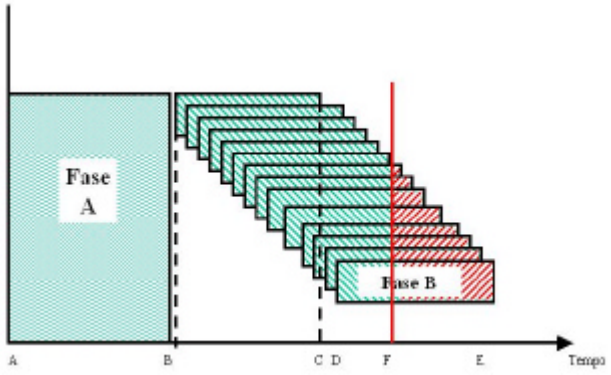


Figura 3. Representação da abordagem dinâmica

2.5 Metodologia de identificação da origem da falha

Uma vez detectada a ocorrência de uma falha de sensor, iniciam-se os procedimentos para a sua correta identificação. A metodologia aplicada na identificação do sensor em falha utiliza o "índice de validade dos sensores" – SVI-D, descrito por Dunia et al (1996). O seu cálculo é baseado na quantificação da participação de cada sensor no erro total de predição. Quanto maior a participação de um sensor no erro total, menor a sua confiabilidade.

2.6 Metodologia de classificação hierárquica

Classificadores hierárquicos buscam separar, ou classificar, amostras contidas em um vasto conjunto de dados. A separação é promovida através da formação de agrupamentos cujos membros apresentam elevada similaridade entre si. Ou seja, objetos similares são alocados em um mesmo grupo, enquanto objetos "não similares" são posicionados em agrupamentos distintos. No presente trabalho aplicou-se classificador hierárquico para separar, de um conjunto total de amostras temporais onde existe a ocorrência de falha, dois grandes conjuntos: com e sem presença de anomalia. Ao promover a classificação torna-se possível definir com precisão o momento da ocorrência de falha operacional.

A similaridade entre dois pontos pode ser entendida como sendo inversamente proporcional a distância espacial entre esses pontos. Em um espaço vetorial R^2 tradicionalmente aplica-se a métrica Euclidiana para quantificar a distância entre o ponto "i" e o ponto "j" no plano XY. Esta métrica pode ser generalizada para um espaço vetorial genérico de dimensão "n" (R^n) através da equação 5.

$$d_{ij} = \left[\sum_{k=1}^{n} \left(x_{i,k} - x_{jk} \right)^2 \right]^{1/2} \tag{5}$$

Apesar de amplamente utilizada, esta não é a única definição possível para a distância entre dois pontos. As métricas conhecidas como City Block (equacao 6) e Minkowski (equação 7), apresentadas por Duda e Hart (2000) muitas vezes apresentam maior eficácia na caracterização de similaridade entre amostras. As métricas City Block e Euclidiana são casos particulares de Minkowski assumindo "r" igual a 1 e 2 respectivamente.

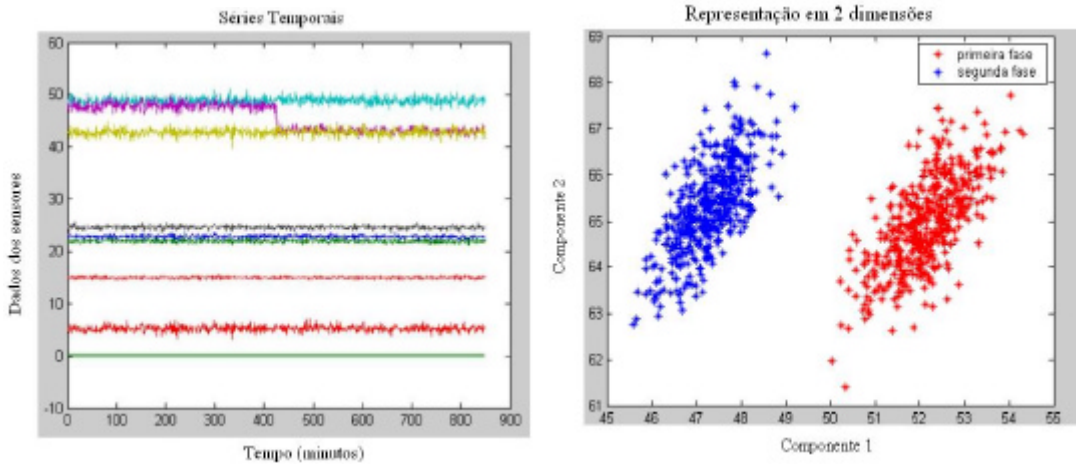
$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \tag{6}$$

$$d_{ij} = \left(\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r} \tag{7}$$

Uma vez definida a métrica o procedimento de classificação em si é bastante simples. Elabora-se uma matriz de distâncias entres todos as observações disponíveis e busca-se o par de pontos com menor afastamento entre si. Ambas as amostras são então unidas, dando origem a um novo ponto. Uma nova matriz de distâncias é então elaborada, selecionando-se novamente o par de menor afastamento. O procedimento prossegue até que a população esteja totalmente alocada em dois grandes agrupamentos.

3. RESULTADOS

Aplicando inicialmente a técnica de PCA como instrumento de visualização dos dados observa-se que esta permite diferenciar os instantes temporais anteriores e posteriores a ocorrência de uma falha. A Figura 4(a) representa a ocorrência de falha em um sensor no instante 400 minutos, de uma série temporal composta por 850 instantes de tempo. A projeção dos dados em um subespaço vetorial de dimensão 2 resulta na separação em dois grandes agrupamentos (Figura 4(b)).



(a) (b)

Figura 4. Representação em R^2 de uma situação de falha de sensor: (a) Séries Temporais com falha no instante 400 min; (b) Projeção dos dados em subespaço vetorial bidimensional.

3.1 Detecção de falha em sensor via Q e SPE

Nesta abordagem, assume-se a fase A com duração de 250 instantes de tempo, estendendo-se do instante inicial 0 ao instante final 250, enquanto a fase B inicia-se no instante 251, prosseguindo até o instante 800. Falha em um dos sensores foi então implementada no instante 301 minutos, ou seja, 50 instantes de tempo após o início da fase B. A matriz da fase A possui dimensão 250x16: 250 instantes de tempo e 16 sensores. Ordenando-se os autovetores do maior autovalor para o menor, foram selecionados os 8 primeiros vetores, denominados "componentes principais". Com esse número de autovetores atinge-se um grau de explicação relativo à variabilidade dos dados superior a 99%.

Conforme apresentado, um dos parâmetros matemáticos de ajuste do método SPE é o valor adotado para l. A partir dos testes efetuados, conclui-se que l's de valor reduzido contribuem para um maior nível de detecção de falhas. Ou seja, elevar a influência do passado sobre o valor presente de SPE incrementa a capacidade de detecção, optando-se neste trabalho por fixar o valor de l em 0,25.

Tabela 2. Percentagem de falhas detectadas em cada abordagem

Severidade (%)	Instante da falha (min)	Q	SPE (lambda 0.25)	Sdist (estático)	Sdist (dinâmico)
2	301			100 [A]	100
2	451			0 [B]	100
2	550			100 [C]	100
2	700			31,25 [D]	18,75
5	301	75,00	81,25	100 [A]	100
5	451	37,75	68,75	62,5 [B]	100
5	550			100 [C]	100
5	700	31,25	37,50	93,75 [D]	93,75
10	301	87,50	87,50	100 [A]	100
10	451	75,00	93,75	93,75 [B]	100
10	550			100 [C]	100
10	700	62,50	93,75	100 [D]	93,75
20	301	100	100	100 [A]	100
20	451	100	100	100 [B]	100
20	550			100 [C]	100
20	700	100	100	100 [D]	100

Tabela 3. Resumo das seqüências empregadas

	Fase A (instante inicial / final) (minutos)	Fase B (instante inicial / final) (minutos)	Falha (instante da ocorrência) (minutos)
Seqüência A	0 / 250	251 / 501	301
Seqüência B	0 / 250	251 / 501	451
Seqüência C	0 / 250	500 / 750	550
Seqüência D	0 / 250	500 / 750	700

Os resultados encontrados para detecção de falhas em sensor estão reunidos na Tabela 2 e demonstram que a distância entre o final da fase A e a ocorrência da falha exerce influência direta sobre o grau de detecção. Na Tabela 2, o índice sobrescrito corresponde ao código de seqüências da Tabela 3. Ao afastar o instante da falha do final da A observa-se um menor índice de detecção. Constata-se, também, que o desempenho de SPE com l=0,25 é superior ao encontrado para Q. A Tabela 2 demonstra, ainda, que falhas de maior severidade são mais facilmente detectadas. É possível que fatores como tipo (sensor de pressão ou vazão) e posicionamento do sensor também influenciem na capacidade de detecção.

Os resultados de detecção por sensor estão representados na Figura 5. A detecção mostrou-se especialmente difícil nos sensores de 3 e 10, ambos do tipo sensor de pressão e localizados nos extremos da dutovia. Através dos resultados obtidos, observou-se que para as modalidades de anomalia implementadas a métrica Spca não se mostra sensível. Por outro lado, o parâmetro Sdist responde fortemente.

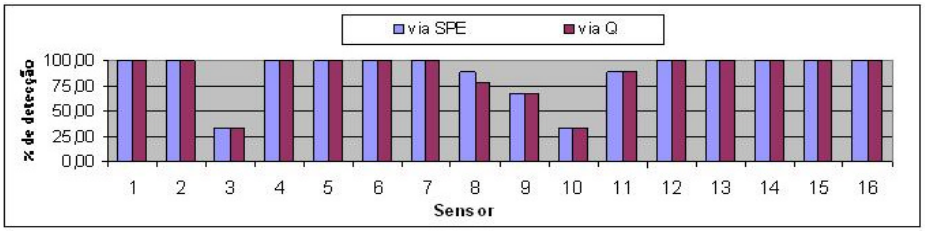


Figura 5. Grau de detecção de falha por sensor

3.2 Detecção de falha em sensor via Sdist e Spca

Os resultados anteriores orientam à concentração deste estudo na métrica Sdist. Devido à presença de ruído nos dados, mostrou-se necessário inicialmente aferir faixa de variação normal do parâmetro Sdist durante operação sem falhas. O valor médio obtido foi de 0,74, com desvio padrão 0,04. Sob tais condições, impondo-se uma margem de aproximadamente 3 desvios padrões em relação à média, considerou-se que valores de Sdist inferiores a 0,6 são indicativos da ocorrência de falhas.

Uma vez estabelecida faixa confiável para Sdist, aplica-se a "abordagem estática" descrita anteriormente. Quatro seqüências de experimento envolvendo falhas de ordens 5%, 10% e 20% na Fase B são implementadas (Tabela 3). Ressalta-se ainda que em virtude do bom desempenho apresentado por esta metodologia optou-se por incorporar também falhas da ordem de 2%.

Para a "seqüência A", obteve-se êxito em todos os teste implementados (Tabela 2). Com a finalidade de aferir se o instante de ocorrência da falha afeta o resultado, implementou-se a "seqüência B". Ao contrario do caso anterior, nestas condições algumas das falhas testadas não puderam ser detectadas. Ao comparar os testes efetuados com falhas nos instantes 301 (A) e 451 (B) constata-se menor índice de detecção na segunda situação. Duas explicações tornam-se possíveis: (i) a posição relativa da falha dentro da fase B, ou (ii) da distância entre o término da fase A e o instante que ocorre a falha.

Na "seqüência C", a distância entre o final da fase A e a ocorrência de falha é maior do que na anterior (Tabela 3), porém todas as falhas testadas foram detectadas. Conclui-se desse modo que a posição relativa da falha dentro da fase B exerce influência mais importante sobre a capacidade de detecção do que a distância absoluta entre o final da fase A e a falha. Finalmente, para a "seqüência D" o resultado encontrado assemelha-se ao da "seqüência B", sugerindo que o desempenho de detecção parece estar mais fortemente relacionado ao instante relativo da falha na fase B do que ao seu afastamento em relação ao final da fase A.

Uma vez avaliado o desempenho da abordagem estática, parte-se para a aproximação "dinâmica". Nesta, definiram-se as seguintes condições de teste: o espaçamento máximo entre o término da fase A e o início da fase B deve ser de 250 instantes de tempo; e valores de Sdist inferiores a 0.6 são considerados indicadores de falha. Com a finalidade de comparar a abordagem dinâmica com a estática, repetiram-se os testes nas condições definidas para as seqüências "A", "B", "C" e "D" (Tabela 3). Em relação às seqüências "A", "B" e "C" a abordagem dinâmica se mostrou totalmente satisfatória, com um índice de detecção de 100% (Tabela 2). Porém, para "seqüência D", a abordagem foi incapaz de detectar inúmeras falhas de menor severidade.

3.3 Identificação do sensor em falha

A partir de SVI-D estabelecem-se os sensores candidatos a estarem apresentando falha. Aplicando-se esta metodologia para falhas ocorridas no instante 300 minutos, tem-se que, para 51% dos casos analisados, a origem da falha foi detectada imediatamente através do primeiro candidato. Ou seja, o sensor apontado como de menor confiabilidade realmente estava em falha. Considerando-se os dois sensores de menor confiabilidade o grau de acerto sobe para 81%. Resultados semelhantes são obtidos para falhas implementadas nos instantes 450 e 700 minutos.

3.4 Classificação hierárquica para falhas em sensor

A abordagem de detecção de falhas via análise de componentes principais consegue atingir elevado grau de eficiência, especialmente através da métrica Sdist. Porém, esta abordagem é incapaz de indicar com precisão o instante inicial da falha. Com a finalidade de superar esta limitação optou-se uma abordagem híbrida, incorporando ferramenta de classificação hierárquica.

Baseando-se na seqüência B da Tabela 3 utilizou-se conjunto de 500 observações, das quais 450 representam situação de operação normal, e as 50 restantes retratam anormalidade. Conforme Figura 6, para falhas com severidades de 10% ou 20%, mais de 90% das observações foram corretamente classificadas. Permitindo assim localizar com elevado grau de precisão o instante inicial de falha. Para situações de menor severidade, a precisão do método também diminui. Melhor desempenho classificatório foi obtido a partir da métrica Minkoswski, com r igual a 4.

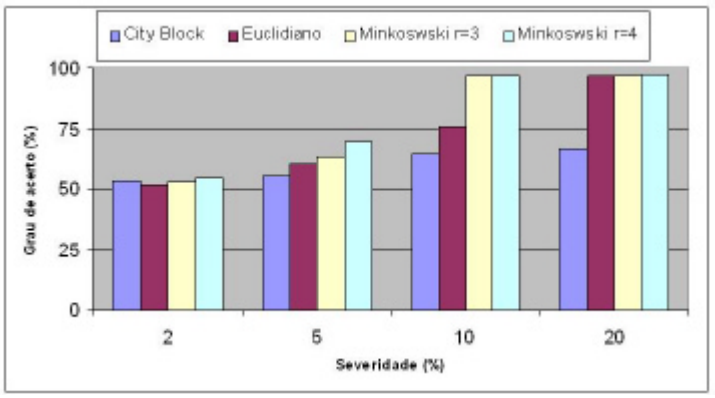


Figura 6. Grau de acerto do classificador hierárquico para falhas de sensor

Os classificadores hierárquicos permitem ainda estimar a severidade da falha a partir do afastamento existente entre os centros dos dois últimos agrupamentos formados. A Figura 7 retrata a relação entre amplitude de falha e afastamento. Deste modo é possível, via classificador hierárquico, estimar o grau de severidade da falha estudada.

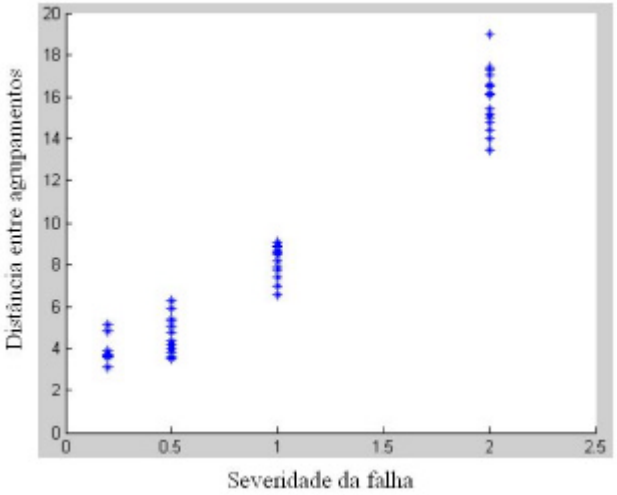


Figura 7. Relação entre a distancia dos agrupamentos e a severidade da falha

3.5 Detecção de vazamento na rede via Q e SPE

De modo similar ao obtido para falhas em sensores, a técnica de PCA aplicada à detecção de vazamentos permitiu a separação em dois agrupamentos dos instantes temporais anteriores e posteriores à ocorrência de vazamento. O resultado é similar ao exibido para falha em sensor na Figura 5. Conforme aplicado para detecção de falha em sensores, a série temporal gerada a partir do simulador de redes de dutos é dividida em duas fases, denominadas "A" e "B". Assume-se a "fase A" com duração de 250 instantes de tempo, estendendo-se do instante inicial ao instante 250. Por outro lado, a "fase B" inicia-se no instante 251, prosseguindo até o instante 800. Vazamento em um dos tubos foi então implementado nos instantes 301, 450, 550 e 700 minutos. Conclui-se igualmente que l's de valor reduzido contribuem para um maior nível de detecção de vazamentos. Ou seja, elevar a influência do passado sobre o valor presente de SPE incrementa a capacidade de detecção tanto de falhas em sensores quanto de vazamentos.

Os resultados encontrados demonstram ainda que o afastamento temporal entre o final da fase A e a ocorrência do vazamento tem influência direta sobre o grau de detecção por SPE (Tabela 4). O grau de detecção para vazamentos ocorridos nos instantes 301 ou 450 é igual ou superior ao alcançado em vazamentos nos instantes 550 ou 700. Verificou-se ainda que o método baseado em SPE com l=0,25 teve desempenho superior à métrica Q. De forma similar ao aferido para severidade da falha, é esperado que o diâmetro de furo também afete a eficiência do sistema. A Tabela 4 confirma tal comportamento: vazamentos com maior diâmetro de furo são mais facilmente detectados.

Adicionalmente, investigou-se o efeito do posicionamento espacial do tubo afetado pelo vazamento na possibilidade deste ser detectado. Novamente, observa-se que anomalias ocorridas na parte central da rede são de mais fácil detecção, quando comparadas as ocorridas em pontos extremos da dutovia.

3.6 Detecção de vazamento na rede via Spca e Sdist

Os resultados encontrados mostram que, para vazamentos assim como verificado em detecção de falhas em sensores, o fator Spca não se mostra sensível, sendo desse modo novamente considerado inadequado. O parâmetro Sdist continuou apresentando elevada sensibilidade.

Inicialmente aplica-se a "abordagem estática", posicionando-se a fase B entre os instantes de tempo 251 e 501, com vazamento ocorrendo a partir do instante 301 – Seqüência A (Tabela 2). Todos os vazamentos analisados para a "seqüência A" geraram valores de Sdist bastante inferiores à faixa de normalidade, promovendo eficiência total na detecção de vazamentos (Tabela 4). O mesmo resultado foi encontrado para vazamentos implementados de acordo com a Seqüência C. Por outro lado, vazamentos baseados nas seqüências B e D foram de mais difícil detecção. Comparando-se o desempenho encontrado para vazamentos com o obtido para falhas em sensores, conclui-se que os dois eventos são altamente similares.

Do mesmo modo que para falhas em sensores, com a finalidade de comparar a abordagem dinâmica com a estática, repetiram-se os testes nas condições impostas pelas seqüências "A", "B", "C" e "D" (Tabela 2). Apenas para "seqüência D" a abordagem dinâmica foi incapaz de detectar 100% dos vazamentos impostos (Tabela 4).

Tabela 4. Percentagem de vazamentos detectados em cada abordagem

Diâmetro de furo (m)	Instante do vazamento (min)	Q	SPE (lambda 0.25)	Sdist (estático)	Sdist (dinâmico)
0,01	301	16,66%	66,66%	100% ^{A}	100%
0,01	451	0%	16,66%	66,67% ^{B}	100%
0,01	550	0%	16,66%	100% ^{C}	100%
0,01	700	0%	0%	50,00% ^{D}	16,66%
0,02	301	100%	100%	100% ^{A}	100%
0,02	451	50%	100%	100% ^{B}	100%
0,02	550	66,66%	83,33%	100% ^{C}	100%
0,02	700	83,33%	83,33%	83,33% ^{D}	83,33%
0,05	301	100%	100%	100% ^{A}	100%
0,05	451	100%	100%	100% ^{B}	100%
0,05	550	100%	100%	100% ^{C}	100%
0,05	700	100%	100%	100% ^{D}	100%

Nota: o índice sobrescrito corresponde ao código de seqüências da Tabela 3.

3.7 Classificador hierárquico para vazamentos

De forma análoga ao observado para falhas em sensores, o classificador hierárquico baseado na distancia de Minkoswski (r = 4) também teve desempenho superior frente situações de vazamentos. A Figura 8 retrata o desempenho dos classificadores para cada severidade de vazamento. Vazamentos com diâmetro de furo de 0.01m não foram bem resolvidos por nenhuma das métricas aqui testadas. Para diâmetros superiores, a distância de Minkoswski teve eficiência superior a 90%. De modo análogo ao observado para falhas em sensores, o classificador também é uma ferramenta útil para estimar severidade de vazamentos.

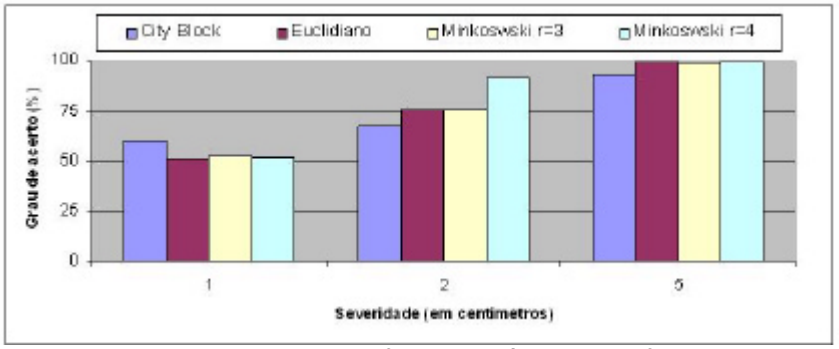


Figura 8. Grau de acerto do classificador hierárquico para falhas de sensor

3.8 Diferenciação entre falhas em sensores e vazamentos

Duas abordagens baseadas em PCA foram testadas com o objetivo de diferenciar ocorrências de falhas em sensores e vazamentos. Na primeira, supõe-se que diferentes tipos de anomalias promovem respostas distintas nas métricas Sdist e Spca (Singhal e Seborg, 2002). Buscou-se então diferenciar anormalidades relativas à falhas em sensores das referentes a vazamentos através da combinação de Spca e Sist. O método, porém, não foi eficaz, pois as falhas em sensores e vazamentos promoveram respostas similares, não permitindo a sua separação.

Uma segunda abordagem baseia-se no efeito da anomalia sobre o parâmetro SVI. Nos cenários de falha de sensor aqui empregados tem-se a ocorrência de um único sensor em falha. Por outro lado, devido as propriedades físicas inerentes a ocorrência de vazamento, este provoca reflexos em diversos sensores localizados nas proximidades do ponto de furo. Desse modo, cenários de vazamentos tendem a se refletir no índice SVI de 2 ou mais sensores. Buscou-se então elaborar metodologia PCA para diferenciação via resposta em SVI, porém, nenhum resultado satisfatório foi encontrado.

Por fim, uma abordagem híbrida baseada em classificação hierárquica e teoria dos protótipos foi aplicada. Protótipos representando falhas em sensores e vazamento são previamente escolhidos. Então, os dados contendo a anomalia em estudo, juntamente com protótipos, são submetidos a procedimento classificatório. O grau de similaridade entre as observações temporais e cada um dos protótipos é capaz de identifica o tipo de anomalia em questão. Essa técnica obteve êxito na diferenciação de falhas em sensores e vazamentos, especialmente nos casos de maior severidade.

4. CONCLUSÃO

Os resultados obtidos confirmam a utilidade do método de PCA na detecção de anomalias em redes de escoamento. Desempenho satisfatório foi obtido tanto na detecção de vazamentos quanto para falhas em sensores. Diferentes abordagens de detecção foram aqui empregadas, tais como cálculo do erro quadrático de previsão e medidas de similaridade. Para ambas as anomalias estudadas, maior grau de detecção foi alcançado a partir do uso da métrica de similaridade Sdist em uma abordagem dinâmica. Elevado nível de sucesso foi obtido quanto à localização de falhas em sensores através de índice baseado em componentes principais. Por outro lado, a abordagem via classificação hierárquica permitiu não apenas estabelecer de modo mais exato o momento de ocorrência de anomalia, como também estimar sua severidade. Ressalta-se ainda o uso da métrica Minkoswski, com r igual a 4 como tendo obtido o maior nível de êxito.

Finalmente, buscou-se diferenciar falhas de sensores de ocorrências de vazamentos. Três metodologias distintas foram aplicadas com tal finalidade. Através de modelos PCA buscou-se inicialmente diferenciar falhas de sensores de vazamentos através dos parâmetros Sdist e Spca. Posteriormente, o parâmetro SVI-D foi usado. Ambas as técnicas não apresentaram êxito. A diferenciação somente foi alcançada através do uso conjunto de teoria dos protótipos e classificação hierárquica.

6. AGRADECIMENTOS

Os autores agradecem ao CNPq, por Bolsas de Doutorado e Pesquisa, e à FINEP, pelo auxílio financeiro.

7. REFERÊNCIAS

DUNIA, S., QIN, J., EDGAR, T., McAVOY, T. Identification of faulty sensors using principal component analysis. AIChE Journal, v.42, n.10, p.2797-2812, 1996.

HART, P., DUDA, R.O. Pattern Classification. 2ª ed. John Wiley Pro., 2000.

SINGHAL, A., SEBORG, S. Pattern Matching in Multivariate Time Series Databases Using a Moving-Window Approach. Ind. Eng. Chem. Res. n.41, p.3822-3838, 2002.

VAZ JUNIOR, C.A. Detecção, localização e quantificação de vazamentos: uma abordagem em séries temporais". Dissertação de Mestrado - Escola de Química da UFRJ, 2006.

WISE, B., GALLAGHER, N. The process chemometrics approach to process monitoring and faulty detection. J. Proc. Cont. v.6, n.6, p.329-348, 1996.

HIERARCHIC CLASSIFICATION AND PCA IN THE DETECTION AND IDENTIFICATION OF FAULTS IN PIPELINE NETWORK

The electronic instrumentation in use increased in the last few decades, with reduction of cost and increase in the number of applications. This progress increases the use of time series for monitoring system and control of the process. Obtaining relevant information from raw data becomes an important field of research. Methodologies such as neural networks, principal components analysis of the main components (PCA), k-means and hierarchic classification, are techniques of data mining with different applications, like clusters analysis and pattern classifications. These tools allow manipulate the large amount of data obtain from chemical and petrochemical plants highly instrumented. In this work, two methodologies were developed to detect and identify imperfections in pipeline network for naphta transport. While the first one is based on hierarchic classification, the second approach uses principal components analysis The hierarchic classification acts in multidimensional data set, and search for excellent partitions or clusters. Hierarchic classification uses established criterion to measure similarity between observations. The final clusters must be homogeneous and well separate, the elements of one class are similar between itself. Metric based in "Euclidean" distance, "City block" and "Minkowski" had been applied. The tool applicable to data classification under "normal" and "abnormal operation" clusters with low errors. The hierarchic classification were also successfully employed in the differentiation between "sensors' faults" and "leaks". The success in the process revealed strongly influenced by the fault severity and by the presence of noise. On the other hand, PCA is based on the existence of high correlation between the pressure sensors and flow meters signals. This correlation is modified during the fault occurrence. PCA models adjusted for data with and without imperfections presented significant distinctions from each other. Metric of similarity, like "Spca" (similarity factor) and "Sdist" (distance similarity factor), as well as indices based on the average squared error were used. Sdist revealed to be efficient in the detention of anomalies of varied severities. Some degree of precision also was achieved for the localization of the sensor in imperfection.

leak detection, principal components analysis, hierarchic classification, time series

Os autores são os únicos responsáveis pelo conteúdo deste artigo.