



RECONCILIAÇÃO ROBUSTA DE DADOS COM SELEÇÃO DE MODELO
SIMULTÂNEA APLICADA AO CÁLCULO DA POTÊNCIA TÉRMICA DE UM
REATOR NUCLEAR TIPO PWR.

Eduardo Damianik Valdetaro da Silva

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Nuclear, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Nuclear.

Orientador: Roberto Schirru

Rio de Janeiro

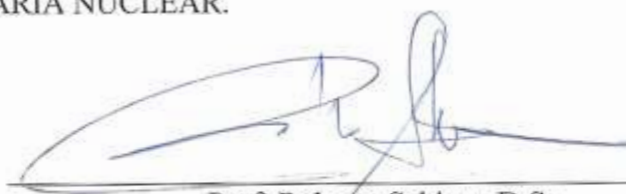
Março de 2012

RECONCILIAÇÃO ROBUSTA DE DADOS COM SELEÇÃO DE MODELO
SIMULTÂNEA APLICADA AO CÁLCULO DA POTÊNCIA TÉRMICA DE UM
REATOR NUCLEAR TIPO PWR.

Eduardo Damianik Valdetaro da Silva

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM
CIÊNCIAS EM ENGENHARIA NUCLEAR.

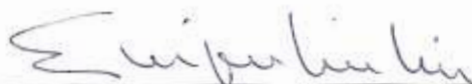
Examinada por:



Prof. Roberto Schirru, D.Sc.



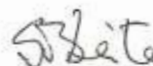
Prof. Eduardo Gomes Dutra do Carmo, D. Sc.



Prof. Enrique Luis Lima, D.Sc.



Prof. Hermes Alves Filho, D.Sc.



Dr. Sérgio de Queiróz Bogado Leite, Ph.D.

RIO DE JANEIRO, RJ – BRASIL

MARÇO DE 2012

Silva, Eduardo Damianik Valdetaro

Reconciliação Robusta de Dados com Seleção de Modelo Simultânea Aplicada ao Cálculo da Potência Térmica de um Reator Nuclear Tipo PWR/ Eduardo Damianik Valdetaro da Silva. – Rio de Janeiro: UFRJ/COPPE, 2012.

XIII, 99 p.: il.; 29,7 cm.

Orientador: Roberto Schirru

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Nuclear, 2012.

Referencias Bibliográficas: p. 94-99.

1. Reconciliação de Dados. 2. Erros Grosseiros 3. Seleção de Modelo 4. Enxame de Partícula (PSO). 5. Critério de Informação de Akaike (AIC). 6. Critério de Informação de Akaike Robusto (AICR). I. Schirru, Roberto. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Nuclear. III. Título.

DEDICATORIAS

*À minha esposa Mônica, aos meus filhos
Júlia e Gabriel.*

AGRADECIMENTOS

Aos amigos Antônio Carlos Alves Lobo, Sergio Ayala e José Carlos Vianna da Rocha pela amizade e pelo interesse, apoio e incentivo demonstrados.

À Ioná Maghali de Oliveira e Andressa dos Santos Nicolau, colegas e alunas de doutorado do Programa de Engenharia Nuclear pelo apoio e incentivo.

Aos engenheiros Edson Prado Azola, Marcelo Sampaio e Décio Brandes, funcionários da Eletrobras Eletronuclear, pelo incentivo e importantes comentários sobre o desempenho térmico e balanço de massa e energia em usinas nucleares e seus aspectos práticos.

Ao Superintendente da Usina de Angra 2, Físico Antônio Carlos Mazzaro, pelo incentivo, interesse e apoio para a realização desta tese de doutorado e à Gerência de Operação de Angra 2, na atual gestão, ao Eng. Fabiano de Almeida Portugal e na gestão anterior, ao Eng. Anselmo Luiz Barbosa de Carvalho e ao Eng. Ricardo Luis Pereira dos Santos.

A todos os amigos, funcionários e professores do Programa de Engenharia Nuclear, particularmente do Laboratório de Monitoração de Processos (LMP/COPPE/PEN) pela atenção, apoio e incentivos recebidos.

Aos funcionários de todas as bibliotecas da UFRJ, particularmente aos funcionários da biblioteca do Centro de Tecnologia, pelo apoio na busca de bibliografias necessárias ao desenvolvimento desta tese. A todos os funcionários da UFRJ e das instituições que mantêm a infra-estrutura que propicia ao aluno a busca de trabalhos e artigos em periódicos em meio eletrônico, pois sem essa possibilidade, o desenvolvimento deste trabalho talvez não fosse possível devido à distância entre a COPPE/UFRJ e o meu local de trabalho.

Ao meu orientador Prof. Roberto Schirru pelo profissionalismo, apoio, incentivo e amizade.

Por fim, à minha família, Mônica minha esposa e aos meus filhos Júlia e Gabriel pelo carinho e apoio demonstrado, pela paciência e tempo dedicado para que eu pudesse realizar esse trabalho. Um agradecimento especial à minha mãe, Neyde Damianik, que entre diversos ensinamentos, me mostrou que a maior riqueza que temos é o conhecimento.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

RECONCILIAÇÃO ROBUSTA DE DADOS COM SELEÇÃO DE MODELO
SIMULTÂNEA APLICADA AO CÁLCULO DA POTÊNCIA TÉRMICA DE UM
REATOR NUCLEAR TIPO PWR.

Eduardo Damianik Valdetaro da Silva

Março / 2012

Orientador: Roberto Schirru

Programa: Engenharia Nuclear

Neste trabalho é desenvolvido um método para a Reconciliação de Dados e Identificação de Erros Grosseiros baseado na aplicação de estatística robusta, utilizando especificamente o estimador de três partes de Hampel e com a capacidade de realizar simultaneamente a seleção de modelo de probabilidade, que corresponde à etapa de ajuste das constantes do estimador robusto. O método proposto utiliza o algoritmo PSO, na sua forma padrão, e pode ser aplicado na monitoração “on-line”, pois usa uma janela de tempo móvel de tamanho determinado. O princípio básico do método é fundamentado na minimização do Critério de Informação de Akaike Robusto (AICR), que é próprio para uso com estimadores robustos. O desenvolvimento teórico indicou a aplicabilidade do método e o mesmo foi testado por meio de simulação em um processo usado em diversos trabalhos com modelos “benchmark” e em caso um exemplo obtido na norma VDI-2048 que fornece diretrizes para aplicação do método clássico de reconciliação de dados em plantas com geração nuclear. Uma aplicação prática com dados reais da Usina Nuclear de Angra 2 é apresentada e corresponde ao cálculo da potência térmica de um reator nuclear tipo PWR. Os resultados simulados e experimentais observados corroboram a proposta do método aqui desenvolvido e possuem desempenhos efetivos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

SIMULTANEOUS ROBUST DATA RECONCILIATION AND MODEL
SELECTION APPLIED TO A PWR NUCLEAR REACTOR POWER
CALCULATION.

Eduardo Damianik Valdetaro da Silva

March / 2012

Advisor: Roberto Schirru

Department: Nuclear Engineering

In this work, a method for Robust Data Reconciliation and Gross Error Identification based on Robust Statistics is developed. The Hampel's three part estimator is used and the method proposed herein is able to simultaneously select a model among a family of possible estimators of the same type, which correspond to automatically adjust the estimator constants. The presented procedure is based on the direct minimization of the Robust Akaike Criteria (AICR), which is the exact counterpart of the Akaike Information Criteria when using Robust Estimators. The standard PSO algorithm is used as a global optimizer and can be used in On-Line monitoring, due to a moving window strategy. The method is tested in a benchmark simulated process used by several authors and in an example from VDI-2048 standard, which give direction for data reconciliation implementation in NPPs. Finally, the proposed procedure is tested in a simplified mass balance with real data from Angra 2 NPP and the results shows that the method is effective.

SUMÁRIO

| | |
|---|----|
| Capítulo 1 – Introdução..... | 1 |
| 1.1 Motivação e Objetivo..... | 1 |
| 1.2 Estrutura..... | 9 |
| Capítulo 2 - Cálculo da Potência Térmica do Reator..... | 11 |
| 2.1 Introdução..... | 11 |
| 2.2 Controle de Carga e Determinação da Potência Térmica do Reator..... | 12 |
| 2.3 Cálculo do Balanço de Massa..... | 15 |
| Capítulo 3 – Reconciliação de Dados e Identificação de Erros Grosseiros..... | 20 |
| 3.1 Introdução..... | 20 |
| 3.2 Reconciliação de Dados Clássica..... | 21 |
| 3.2.1 Determinação da matriz de covariância verdadeira..... | 23 |
| 3.2.2 Intervalo de Confiança..... | 25 |
| 3.2.3 Aplicação da Reconciliação de Dados Clássica..... | 27 |
| 3.2.4 Identificação de Erros Grosseiros..... | 28 |
| 3.2.5 Cálculo do Vetor de Correção e da Matriz de Covariância dos Valores Corrigidos..... | 29 |
| 3.3 Fundamentos e Considerações sobre outros Métodos de Reconciliação de Dados e Identificação de Erros Grosseiros..... | 30 |
| Capítulo 4 – Estatística Robusta e Estimadores Robustos..... | 38 |
| 4.1 Introdução..... | 38 |
| 4.2 Estatística Robusta e Estimadores Robustos..... | 39 |
| 4.3 Estimador Robusto Redescendente de Três Partes de Hampel..... | 46 |
| 4.4 Diferentes Critérios para Identificação de Erros Grosseiros..... | 47 |
| 4.5 Ajuste do Estimador Redescendente de Hampel..... | 49 |

| | |
|--|----|
| 4.6 Considerações sobre os Estimadores Robustos e o Método de Ajuste das Constantes do Estimador Redescendente de Hampel..... | 52 |
| 4.7 Algoritmo de Otimização por Enxame de Partículas..... | 54 |
| | |
| Capítulo 5 – Reconciliação Robusta de Dados com Seleção de Modelo de Probabilidade Simultânea..... | 60 |
| 5.1 Introdução..... | 60 |
| 5.2 Critério de Informação de Akaike Robusto..... | 61 |
| 5.3 Método simultâneo para Reconciliação Robusta de Dados, Identificação de Erros Grosseiros e Seleção de Modelo baseado no Critério de Informação de Akaike Robusto (AICR)..... | 63 |
| 5.4 Considerações sobre o Método de Reconciliação Robusta de Dados com Seleção de Modelo Simultânea (RDSMS)..... | 68 |
| | |
| Capítulo 6 – Resultados..... | 70 |
| 6.1 Introdução..... | 70 |
| 6.2 Exemplo Não linear (PAI e FISHER, 1988)..... | 71 |
| 6.3 Exemplo de Cálculo da Potência do Reator baseado na norma VDI-2048..... | 76 |
| 6.3.1 Considerações sobre o Cálculo da Potência Térmica do Reator..... | 80 |
| 6.4 Cálculo da Potência Térmica do Reator com Dados Reais obtidos na Usina Nuclear de Angra 2..... | 83 |
| 6.4.1 Balanço de Massa Simplificado da Usina de Angra 2..... | 84 |
| | |
| Capítulo 7 – Conclusões..... | 89 |
| 7.1 Introdução..... | 89 |
| 7.2 Conclusões Gerais..... | 89 |
| 7.3 Sugestões para Trabalhos Futuros..... | 92 |

NOMENCLATURA

| Símbolos | Descrição | |
|----------------|---|------------|
| P_t | Potência Térmica do Reator | [MWt] |
| P_{Rt} | Potência Térmica do Reator Reconciliada | [MWt] |
| \dot{Q}_{GV} | Carga térmica do GV | [MWt] |
| m_{ag} | Vazão Total de Água de Alimentação Principal | [Kg/s] |
| h_s | Entalpia do fluido na saída do GV | [Btu/Kg.s] |
| h_e | Entalpia do fluido na entrada do GV | [Btu/Kg.s] |
| m_{GV1} | Vazão de vapor do Gerador de Vapor 1 | [Kg/s] |
| m_{GV2} | Vazão de vapor do Gerador de Vapor 2 | [Kg/s] |
| m_{ag1} | Vazão de água de alimentação 1 | [Kg/s] |
| m_{ag2} | Vazão de água de alimentação 2 | [Kg/s] |
| m_v | Vazão total de vapor | [Kg/s] |
| m_c | Vazão de Condensado | [Kg/s] |
| m_{A7} | Vazão da extração A7 | [Kg/s] |
| m_{A6} | Vazão da extração A6 | [Kg/s] |
| m_{A5} | Vazão da extração A5 | [Kg/s] |
| m_{HPC} | vazão de retorno de condensado de alta pressão | [Kg/s] |
| m_a | Vazão de vapor da turbina de alta para a de baixa | [Kg/s] |
| \mathbf{x}^M | Vetor de variáveis medidas | |
| \mathbf{x} | Vetor de variáveis estimadas | |
| \mathbf{p} | Conjunto de parâmetros do problema | |
| \mathbf{u} | Conjunto das variáveis não medidas | |
| \mathbf{h} | Conjunto de restrições de igualdades, | |
| \mathbf{g} | Conjunto de restrições de desigualdade | |
| S_X | Matriz de covariância estimada | |
| x_i | I-ésima variável medida | |
| n | Número de variáveis medidas | |
| m | Número de amostras de uma determinada variável medida | |
| λ_p | Fator indicativo da probabilidade do intervalo de confiança | |
| σ_{xi} | Desvio padrão da variável x_i | |

| | |
|----------------------|---|
| \mathbf{v} | Vetor de Correção ou de desvios |
| S_v | Matriz de covariância do erro |
| \mathbf{v} | Vetor de correção |
| $\tilde{\mathbf{x}}$ | Vetor de medidas corrigido ou reconciliado. |
| \bar{x}_i | Valor esperado da variável i . |
| h | Tamanho da janela de tempo em amostras |
| γ | Sensibilidade a erros grosseiros |
| λ | Sensibilidade em relação a desvios da medida |
| ζ | Ponto de rejeição |
| ε | ponto de ruptura do estimador |
| a, b e c | Constantes de ajuste do estimador de três partes de Hampel |
| n_o | Número de erros grosseiros |
| P_i | Vetor indicando a melhor posição individual (PSO) |
| P_g | Vetor indicando a melhor posição global (PSO) |
| w | Fator de inércia do PSO |
| c_1 | Constante que ajusta o peso relativo a componente do indivíduo no PSO |
| c_2 | Constante que ajusta o peso relativo a componente do grupo no PSO |
| η | Componente relativa ao ruído aleatório |
| ι | Componente relativa ao Erro Grosseiro |
| X_e | Vetor indicando valores exatos da solução |
| X_{opt} | Vetor indicando valores calculados pelo método RDSMS |

| Siglas | Descrição (Descrição Original) |
|---------------|---|
| AE | Análise Exploratória |
| AIC | Critério de Informação de Akaike (Akaike Information Criteria) |
| AICR | Critério de Informação de Akaike Robusto (Robust Akaike Information Criteria) |
| EA | Erros Aleatórios |
| EG | Erros Grosseiros |
| ES | Erros Sistemáticos |
| GA | Algoritmo Genético (Genetic Algorithm) |
| GV | Gerador de Vapor (Steam Generator) |
| HTPRE | Estimador Redescendente de Três Partes de Hampel (Hampel's Three Part Redescendent Estimator) |
| IEG | Identificação de Erros Grosseiros |
| IF | Função de Influência (Influence Function) |
| LSE | Estimador de Mínimos Quadrados (Least Square Estimator) |
| MAD | Desvio Absoluto da Mediana (Median Absolute Deviation) |
| mGA | Algoritmo Genéticos modificado (Genetic Algorithm) |
| MLE | Estimador de Máxima Verossimilhança (Maximum Likelihood Estimator) |
| MLR | Retificação por Máxima Verossimilhança (Maximum Likelihood Rectification) |
| MS | Seleção de Modelo de Probabilidade (Model Selection) |
| NPP | Usina Nuclear (Nuclear Power Plant) |
| PSO | Algoritmo de Enxame de Partículas (Particle Swarm Algorithm) |
| PWR | Reator à Água Pressurizada (Pressurized Water Reactor) |
| RD | Reconciliação de Dados |
| RDC | Método de Reconciliação de Dados Clássico |
| RDSMS | Reconciliação Robusta de Dados com Seleção de Modelo Simultânea |
| RRD | Reconciliação Robusta de Dados |
| SQP | Programação Quadrática Sucessiva (Successive Quadratic Program) |
| VA | Variável Aleatória |
| WLS | Mínimos Quadrados Ponderados (Weighted Least Square) |

CAPÍTULO 1:

INTRODUÇÃO

1.1 – Motivação e Objetivo

A utilização da técnica de reconciliação de dados (RD) e de Identificação de Erros Grosseiros (IEG) na indústria química e petroquímica é considerada como um ponto importante na otimização e monitoração de processos. A mesma continua em desenvolvimento desde que a técnica foi apresentada por KUEHN e DAVIDSON (1961), os quais são considerados como os precursores da aplicação do método de RD (MORAD *et al.*, 2005).

A aplicação da técnica de Reconciliação de Dados permite reduzir a margem de erro na medição das variáveis e parâmetros do processo, o que melhora a visualização do comportamento da planta e com uma medição mais precisa permite uma melhor tomada de decisão (PRATA, 2009).

Em Usinas Nucleares o interesse pela técnica de RD tem aumentado devido à necessidade de se manter o balanço de massa e energia sob rigoroso controle e acompanhamento. O balanço de massa e de energia é um conjunto de equações elaborado a partir das leis que regem os fenômenos físicos e químicos e descrevem o comportamento processo e o seu controle e acompanhamento permite obter diversos benefícios, como por exemplo, a melhora na tomada de decisão e a realização de uma operação mais segura. Além desses benefícios podemos citar o retorno financeiro direto ou indireto, que advém com o aumento da precisão da medida e diminuição da incerteza na medição do estado da planta proporcionada pelos métodos de Reconciliação de Dados e Identificação de Erros Grosseiros.

Na área nuclear, trabalhos de GRAUF *et al.* (2000), STREIT *et al.* (2005), JANSKY (2006, 2007), AZOLA *et al.* (2009) e VALDETARO e SCHIRRU (2011a,

2011b) indicam que as técnicas de RD e IEG aplicadas à indústria nuclear podem trazer resultados concretos e de interesse prático.

Notadamente, a determinação da potência térmica do reator em uma Usina Nuclear de Potência (NPP) é um assunto de importância e utiliza os balanços de massa e de energia da planta. Seu cálculo é baseado na medição de vazão, cujas variáveis medidas possuem erros de medição inerentes à instrumentação utilizada e podem chegar a vários níveis percentuais (ANDRADE *et al.*, 2002).

Devido a condições determinadas pelo órgão regulador, uma Usina Nuclear só pode operar dentro de limites pré-estabelecidos, que no caso da potência térmica do reator, é de 100% com uma faixa de 2% de tolerância. A operação dentro desses limites deve-se ao fato que no projeto da planta, foi feita uma análise do resfriamento de emergência do núcleo do reator dentro dos limites mencionados (STREIT *et al.*, 2005).

Dessa forma, a utilização da técnica de reconciliação de dados e a identificação de erros grosseiros parecem indicar um caminho, onde a diminuição da incerteza ou aumento na exatidão e na precisão da medida permitirá uma margem maior e segura para a operação da usina com um nível de potência maior e dentro dos patamares de segurança exigidos (STREIT *et al.*, 2005).

Durante a monitoração da planta, as variáveis de processo podem ser corrompidas por erros aleatórios (EA) ou erros grosseiros (EG), também denominados Erros Sistemáticos (ES). Essas incertezas na medida podem afetar a operação segura do processo, bem como, impedir o fechamento dos balanços de massa e energia ou afetar a qualidade dos dados utilizados por outras plantas, além de afetar o cálculo de desempenho do sistema (SODERSTROM *et al.*, 2000).

As medidas contaminadas por erros grosseiros ou erros sistemáticos que não satisfazem as restrições do processo necessitam ser reconciliadas, ou seja, os erros grosseiros devem ser eliminados ou substituídos e os valores medidos devem então ser corrigidos a fim de satisfazer as restrições relativas ao processo de forma a minimizar o erro quadrático (TJOA e BIEGLER, 1991). A técnica de Reconciliação de Dados e Estimção de Parâmetros é a técnica que realiza esses ajustes e permite que se utilizem os valores reconciliados como dados ou parâmetros do processo.

A técnica padrão ou clássica de Reconciliação de Dados baseia-se na determinação de vetor de correção para as medidas de processo a partir da minimização do erro quadrático, sujeito as restrições das equações do processo. Nesse caso, pressupõe-se que as variáveis estão contaminadas com um ruído que possui uma distribuição de probabilidade conhecida (p. ex., distribuição Normal). A partir da estimativa da matriz de covariância das variáveis medidas do processo e de sua média, pode-se estimar o vetor de correção para as medidas do processo. Após a aplicação dessa correção, os valores reconciliados são então determinados.

A Identificação de Erros Grosseiros (IEG) é feita por meio de testes estatísticos (TJOA e BIEGLER, 1991) e no método clássico é comum utilizar a matriz de covariância do erro e testes de chi-quadrado para determinar um patamar que identifique medições errôneas, as quais devem ser eliminadas antes de se aplicar a técnica de RD (VDI-2048, 2000), o que torna o processo de reconciliação de dados um procedimento iterativo (TJOA e BIEGLER, 1991).

Na formulação clássica do problema da reconciliação de dados, quando a medida do erro possui função de distribuição Normal, com média nula e variância conhecida, a função objetivo corresponde ao estimador de Máxima Verossimilhança (MLE) das medidas do processo (ARORA e BIEGLER, 2001).

Os valores das medidas do processo podem estar contaminados por erros aleatórios ou possivelmente por erros sistemáticos ou erros grosseiros. Erros aleatórios (EA) são aqueles que se manifestam como uma influência incontrollável e não tendenciosa ou média nula (VDI-2048, 2000). Erros Grosseiros (EG) ou sistemáticos (ES) refletem tipicamente no processo como um desvio na medida ou uma perturbação no processo, que pode atingir uma ou mais variáveis afetando negativamente os valores reconciliados (ARORA e BIEGLER, 2001; SODERSTROM *et al.*, 2000; MEI *et al.*, 2007).

A complexidade do problema da Reconciliação de Dados aumenta à medida que erros grosseiros ou sistemáticos estão presentes na medida proveniente do processo e causam uma estimação incorreta do estado da planta (ARORA e BIEGLER, 2001).

Essa influência adversa devido à presença de erros grosseiros acaba afetando o Estimador de Máxima Verossimilhança (MLE) devido a sua falta de robustez, que no caso da RD clássica é o estimador de Mínimos Quadrados Ponderados (WLS).

A robustez de um estimador está relacionada com a capacidade de um estimador em lidar com pequenos desvios do modelo probabilístico real e seus desdobramentos e conseqüências e proporcionar estimativas estáveis e confiáveis.

No caso da RD clássica, a falta de robustez do estimador está relacionada com a possibilidade de que apenas um erro grosseiro possa causar um desvio significativo, desde que a magnitude desse desvio seja grande o suficiente. Uma forma de lidar com esse problema é a utilização de outros tipos de estimadores, que são mais apropriados para lidar com erros grosseiros (ROUSSEEUW e LEROY, 1987).

Devido à ausência de robustez de alguns estimadores a remoção prévia de erros grosseiros é uma parte fundamental no processo de reconciliação de dados. A presença de erros aleatórios ou mesmo sistemáticos nas medidas do processo fazem com que não seja possível fechar os balanços de massa e de energia, os quais são representados por uma série de equações ou inequações de restrições que descrevem as relações fundamentais do processo, como as relações de equilíbrio do sistema (OZYURT e PIKE, 2004). Dessa forma, as técnicas de reconciliação de dados e Identificação de Erros grosseiros devem ser aplicadas a fim de remover esses erros, e que as medidas corrigidas ou valores reconciliados, que satisfazem as restrições de processo, possam ser usadas na modelagem de parâmetros do processo.

Extensos estudos já foram e tem sido realizados com o intuito de eliminar ou mitigar a influência de erros aleatórios e principalmente erros grosseiros. Métodos baseados em multiplicadores de Lagrange ou Multiplicadores de Lagrange combinados com matriz de projeção aplicados à reconciliação de dados em estado permanente foram desenvolvidos, como os apresentados em CROWE *et al.* (1983) e SERTH e HEENAN (1986).

PAI e FISHER (1988) propõem o uso do método de Broyden a fim de evitar repetidas vezes o calculo do Jacobiano. TJOA e BIEGLER (1991) utilizaram programação quadrática adaptada para aproveitar a estrutura da função objetivo, uma

função distribuição bivariada. O erro grosseiro foi detectado por meio de testes estatísticos baseado na função de distribuição combinada, o que permitiu eliminar o processo iterativo de remoção de erros grosseiros e em seguida aplicar o método de reconciliação de dados (TJOA e BIEGLER, 1991; OZYURT e PIKE, 2004).

Outras técnicas para a realização da reconciliação de dados simultaneamente com a identificação de erros grosseiros foram propostas como o método apresentado por YAMAMURA *et al.* (1988) para a detecção de erros grosseiros baseado no Critério de Informação de Akaike (AIC) e em uma estrutura com base em Mínimos Quadrados. O método de “Branch and Bound” foi utilizado para resolução do problema. A estratégia pode ser automatizada como sugerido por SODERSTROM *et al.* (2000) usando programação inteira-mista (ARORA e BIEGLER, 2001).

A técnica citada acima minimiza uma função objetivo modificada e adiciona novas restrições de forma a considerar a detecção de erros grosseiros implícita na estratégia de reconciliação de dados. Dada a natureza combinatória do método, a sua “performance” ou desempenho é melhor quando o número de variáveis é pequeno SODERSTROM *et al.* (2000).

Em paralelo com diversos outros métodos, JOHNSTON e KRAMER (1995) introduziram uma nova abordagem para a reconciliação de dados e identificação de erros grosseiros denominada “Maximum Likelihood Rectification” ou Retificação por Máxima Verossimilhança (MLR), a qual é baseada nos trabalhos prévios de HUBER (1981) e HAMPEL (1974) relacionados com Estimção Robusta, Regressão Robusta e Função de Influência (IF). A abordagem proposta tem a vantagem de não dividir as medidas em diferentes classes como alguns dos métodos previamente apresentados e eliminou também a característica combinatória de outros métodos, além de não necessitar qualquer procedimento iterativo para a detecção e eliminação de erros grosseiros, corrigindo-os simultaneamente com o processo de reconciliação de dados (JOHNSTON e KRAMER, 1995).

Diversos outros métodos baseados em Estatística Robusta foram posteriormente desenvolvidos, ALBUQUERQUE e BIEGLER (1996) usam como função objetivo a função Fair (Fair Function), um estimador robusto, que de acordo com suas propriedades matemáticas pode reduzir os efeitos causados por Erros Grosseiros. A

presença de erros sistemáticos ou grosseiros é determinada por meio de Análise Exploratória (AE), que utiliza, por exemplo, medianas, “box plots”, quartis e outros. Uma vez que a presença de erro grosseiro foi determinada, a sua confirmação pode ser obtida eliminando-se a medida e resolvendo o problema utilizando Mínimos Quadrados.

ARORA e BIEGLER (2001) compararam o estimador robusto de três partes de Hampel com o M-estimador, a função Fair, e concluíram que o primeiro é mais robusto e possui um ponto de corte que permite a aplicação do método simultaneamente com a reconciliação de dados, dispensando ainda o uso de Análise Exploratória (ARORA e BIEGLER, 2001). De forma a ajustar as constantes do estimador redescendente de três partes de Hampel (HTPRE), um método de dois passos foi desenvolvido. Esse ajuste procura ajustar o modelo de probabilidade aos dados adquiridos do processo e é baseado na minimização do Critério de Informação de Akaike (AIC).

Apesar da eficiência do Estimador Redescendente apresentado em ARORA e BIEGLER (2001), um importante aspecto é o ajuste das constantes dos estimadores robustos, especialmente os Estimadores Redescendentes, onde as constantes do modelo de probabilidade devem ser ajustadas aos dados estatísticos a fim de que o mesmo estime o estado da planta de forma correta e balanceada (HAMPEL *et al.*, 1986).

O ajuste das constantes do estimador ou seleção de modelo (MS) pode ser feita de diversas formas, mas o princípio é minimizar o M-estimador, que é função do erro, a diferença entre o valor medido e o valor estimado, em relação às constantes do M-estimador. Os M-estimadores em geral são obtidos usando o princípio da Máxima Verossimilhança (Maximum Likelihood) e é da aplicação desse princípio que advém a letra M de M-estimadores.

Métodos baseados na estatística robusta têm a vantagem de diminuir a influência do erro grosseiro, sem a necessidade de utilizar os recursos da análise exploratória (AE). Por outro lado, quando a função objetivo é um estimador robusto, a mesma pode ser não linear e não convexa, o que traz a seguinte dificuldade: a solução obtida com o algoritmo de otimização pode ser um mínimo local, dessa forma, pode ser necessário recorrer a um método de otimização global (ARORA e BIEGLER, 2001).

Assim, é possível visualizar que métodos de otimização baseados na teoria evolucionária, como Algoritmos Genéticos (GA), podem ser um caminho promissor quando aplicados à Reconciliação Robusta de Dados (RRD) e Identificação de Erros Grosseiros devido a suas características simples, que é o uso apenas de equações algébricas e a ausência do cálculo do Jacobiano, como nos métodos precursores.

Podemos citar alguns exemplos extraídos da literatura do uso do GA: MOROS *et al.* (1996) usaram o GA para gerar parâmetros iniciais para um modelo cinético de um processo catalítico e WONGRAT *et al.* (2005) aplicaram com sucesso um algoritmo genético modificado (mGA) proposto por WASANAPRADIT (2000) e aplicado à Reconciliação de Dados Não Linear.

WONGRAT *et al.* (2005) associaram a robustez dos M-estimadores e a relativa simplicidade do algoritmo genético (GA) e propuseram um algoritmo genético modificado (mGA). Esse algoritmo apresentou bom desempenho e utilizou o Estimador Redescendente de Três Partes de Hampel (HTPRE). O ajuste das constantes do estimador redescendente foi realizado em dois passos distintos, cujo procedimento foi minimizar o Critério de Informação de Akaike, de forma semelhante ao que foi apresentado em ARORA e BIEGLER (2001).

Apesar da efetiva aplicação de um algoritmo evolucionário (mGA) à solução do problema de Reconciliação de Dados e Identificação de Erros Grosseiros e da ausência de cálculos complexos, o método pode levar um tempo longo nas simulações. Na perspectiva de diminuir o tempo de cálculo, VALDETARO e SCHIRRU (2009) propuseram o uso do algoritmo por enxame de partículas (PSO) em substituição ao algoritmo genético e utilizaram também o Estimador Redescendente de Três Partes de Hampel, onde o ajuste das constantes do estimador foi realizado de forma semelhante ao apresentado em WONGRAT *et al.* (2005) e ARORA e BIEGLER (2001). O ajuste foi realizado em dois passos. O primeiro para ajustar o estimador robusto e o outro passo para solucionar o problema de reconciliação de dados simultaneamente com a Identificação de Erros Grosseiros.

Em VALDETARO e SCHIRRU (2009), os resultados também foram efetivos, indicando um processamento mais rápido do que o algoritmo genético modificado e o foco desse trabalho foi desenvolver um método para posteriormente aplicá-lo ao cálculo

da potência térmica de um reator nuclear. Os resultados obtidos foram semelhantes aos exemplos propostos em WONGRAT *et al.* (2005) e PAI e FISHER (1988).

No mesmo período, PRATA (2009) apresentou em sua tese de doutorado o uso do algoritmo baseado em enxame de partículas (PSO) aplicado a reconciliação de dados robusta e dinâmica utilizando o estimador de Welsch e em PRATA *et al.* (2009) a aplicação do PSO na reconciliação de dados robusta e dinâmica, com função objetivo quadrática, no senso de mínimos quadrados, em um reator industrial de produção de polipropileno. Os resultados mostraram que o método é confiável e efetivo, mesmo num cenário de monitoração “on-line” e de tempo real.

No trabalho de VALDETARO e SCHIRRU (2009), apesar do bom desempenho obtido com o estimador redescendente de três partes de Hampel (HTPRE) na aplicação do método de reconciliação de dados e Identificação de Erros Grosseiros em conjunto com o algoritmo de enxame de partículas (PSO), o ajuste das constantes do estimador redescendente foi feita em dois passos distintos e em separado do processo de reconciliação de dados. Esse processo em separado e com duas fases consome muito tempo e é necessária a realização de duas etapas de minimização para o correto ajuste do estimador e cálculo da RD e IEG.

Por outro lado, verificamos que o uso de um critério de informação poderia vir a ser um meio efetivo para a escolha de um modelo que se ajustasse a um conjunto de dados (HAMPEL *et al.*, 1986), visto que a primeira etapa indicada acima é a de ajuste das constantes do estimador e baseada no Critério de Informação de Akaike. Entretanto, o critério de informação de Akaike não é próprio para o uso com estimadores robustos e as outras formas de ajustes das constantes de estimadores robustos que não utilizam critérios de informação, normalmente, fazem uso de simulação e do Método de Monte Carlo para obter os valores dessas constantes e das eficiências relativas dos estimadores robustos (OZYURT e PIKE, 2004). Em alguns estimadores (p. ex., função Fair) o ajuste se dá pela eficiência assintótica, mas são aproximações pouco precisas (OZYURT e PIKE, 2004), o que em ambos os casos pode não ser eficiente.

Assim, neste trabalho, o objetivo foi desenvolver um novo método que simultaneamente realizasse a Reconciliação de Dados, a Identificação de Erros Grosseiros e a Seleção do Modelo, ou seja, o ajuste das constantes do estimador

robusto. O método proposto de Reconciliação Robusta de Dados com Seleção de Modelo Simultânea (RDSMS) é baseado na minimização do Critério de Informação de Akaike Robusto (AICR), o qual foi proposto por RONCHETTI (1985, 1997a) e é próprio para uso com M-estimadores. O uso do AICR permite eliminar as etapas em separado de ajuste das constantes do estimador robusto como apresentado em métodos predecessores e permite fazer a sintonia do estimador robusto baseado em um critério direto.

O método simultâneo para RD, IEG e MS utiliza o estimador de três partes de Hampel na parte relativa ao ajuste do Critério de Informação de Akaike Robusto na função objetivo com modificações para lidar com os limites de validade do estimador e o algoritmo de enxame de partículas (Particle Swarm Optimization ou PSO) para atuar como algoritmo de otimização global.

O método foi aplicado a três exemplos, dois simulados e outro com dados reais de processo. Dois dos exemplos foram retirados da literatura (VDI-2048, 2000; WONGRAT *et al*, 2005) e o terceiro exemplo corresponde à aplicação do cálculo da potência térmica do reator da Usina Nuclear de Angra 2.

1.2 - Estrutura

Esse trabalho está organizado de acordo com os tópicos apresentados a seguir:

Capítulo 2: Nesse capítulo será apresentado um sistema simplificado do circuito secundário de uma usina nuclear do tipo PWR cujo modelo foi extraído da norma VDI-2048 (2000). O objetivo é fornecer uma visão geral, embora simplificada, sobre o cálculo da potência térmica de um reator nuclear típico e serão apresentadas as equações necessárias para a realização de um balanço de massa da planta a fim de se determinar a vazão de água de alimentação corrigida e conseqüentemente a potência térmica do reator. Aqui é ressaltada a importância de se ter os balanços de massa e energia sob estrito controle e exemplificado os efeitos da aplicação do método de Reconciliação de Dados.

Capítulo 3: Será apresentado o problema da Reconciliação de Dados e de Identificação de Erros Grosseiros com uma revisão bibliográfica básica com o intuito de apresentar os principais métodos de reconciliação de dados e embasar a formulação do ajuste automático proposto nessa tese. Consta desse capítulo a apresentação da solução clássica do problema de RD e IEG em regime estacionário e a importância na determinação da matriz de covariância do processo e do problema de otimização envolvido na RD.

Capítulo 4: Nesse capítulo será apresentada uma visão sucinta sobre estatística robusta, em especial sobre os estimadores. Em seguida será apresentado o estimador robusto de três partes de Hampel na resolução do problema de Reconciliação de Dados. Na terceira parte, será apresentado o método de ajuste do estimador de três partes de Hampel, baseado Critério de Informação de Akaike. Na última seção será apresentado o método de otimização por Enxame de Partículas (PSO), que devido as suas características é utilizado como um otimizador global e aplicado aos exemplos deste trabalho.

Capítulo 5: Nesse capítulo será apresentado o método para Reconciliação Robusta de Dados e Seleção de Modelo Simultânea (RDSMS) com utilização do algoritmo de otimização por Enxame de Partículas (PSO).

Capítulo 6: Nesse capítulo serão apresentados os resultados obtidos

Capítulo 7: Nesse capítulo serão apresentadas as conclusões finais e as considerações sobre a contribuição original proposta neste trabalho e sugestões para trabalhos futuros.

CAPÍTULO 2:

Cálculo da Potência Térmica do Reator

2.1 - Introdução

O objetivo deste capítulo é apresentar o método de cálculo da potência térmica de um reator nuclear de uma usina do tipo PWR, como por exemplo, a Usina Nuclear de Angra 2. O método apresentado para o cálculo da potência térmica do reator contém simplificações, mas representa a forma de cálculo utilizada em um cenário real.

Será apresentada também uma visão sucinta do controle de carga e da utilização da técnica de Reconciliação de Dados, que pode ser um instrumento eficiente para o cálculo da potência térmica de um reator (P_t), permitindo uma maior exatidão no cálculo dessa medida.

Ao se aplicar a técnica de RD, obtêm-se uma maior exatidão no cálculo da potência térmica do reator (P_t), dessa forma, pode-se operar o reator dentro de parâmetros de segurança mais estritos. Caso o valor da Potência Térmica Reconciliada (P_{Rt}) esteja abaixo da potência nominal da usina (100%), pode-se elevar a mesma até alcançar o limite de operação permitido e então se beneficiar dessa margem de segurança maior, inclusive com retorno financeiro proveniente de uma maior geração de energia (STREIT *et al.*, 2005).

Como o cálculo da P_t depende diretamente do cálculo da vazão de água de alimentação, na seção 2.3 serão apresentadas as equações de balanço de massa de um circuito secundário simplificado de uma usina nuclear do tipo PWR, o qual foi baseado na norma VDI-2048 (2000) e será utilizado como base nos problemas-teste deste trabalho.

2.2 – Controle de Carga e Determinação da Potência Térmica do Reator

As usinas nucleares do tipo PWR possuem dois circuitos básicos, representados sucintamente na figura 2.1, que são o circuito primário, por onde passa o fluido refrigerante do reator e que circula pela parte interna dos tubos em U localizados dentro dos Geradores de Vapor (GV) (figura 2.1, linha em negrito) e o circuito secundário ou circuito água-vapor, que flui pelo lado do casco do GV (figura 2.1, linha contínua) - (AZOLA *et al.*, 2009). O circuito terciário ou fonte fria foi omitido por efeito de simplificação.

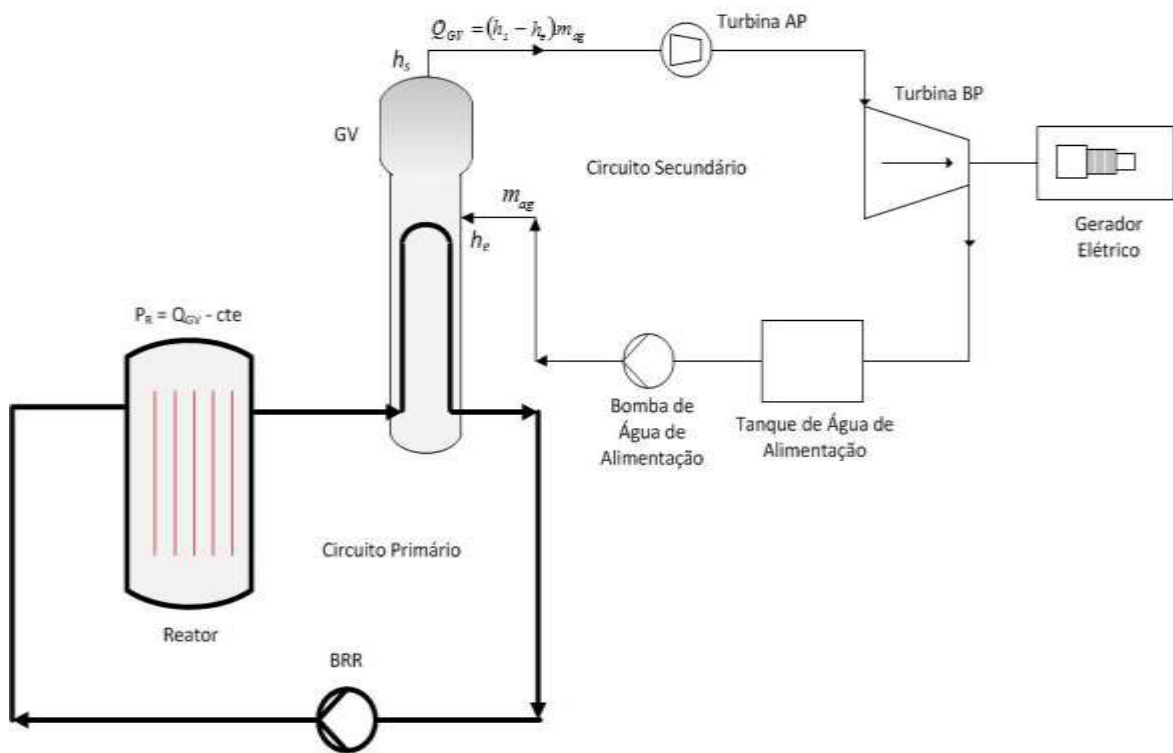


Figura 2.1: – Diagrama Simplificado de uma usina nuclear tipo PWR (circuito primário e secundário)

Uma malha de controle de potência elétrica é utilizada para o controle de carga da planta. A potência medida é comparada com um valor limite e o ajuste necessário é aplicado ao elemento final de controle, que são as válvulas de controle de vazão de vapor que alimentam o estágio de alta pressão da turbina a vapor (AZOLA *et al.*, 2009).

A potência térmica do reator não deve ultrapassar 100% da carga nominal e caso a potência térmica ultrapasse 102%, sistemas de limitação e proteção do reator atuam, impedindo a ultrapassagem desse limite, que é definido pela análise de segurança do reator.

O cálculo da potência térmica do reator pode ser obtido determinando-se a carga térmica do Gerador de Vapor (\dot{Q}_{GV}), a qual é diretamente proporcional ao valor da vazão de água de alimentação principal (m_{ag}), deduzindo-se desse valor as cargas térmicas relativas ao circuito primário (AZOLA *et al.*, 2009), as quais não sofrem variações significativas, podendo ser consideradas contantes. O cálculo da P_t está indicado na equação abaixo.

$$P_t = \dot{Q}_{GV} - cte \quad (2.1)$$

A carga térmica do Gerador de Vapor (\dot{Q}_{GV}) é calculada pela equação 2.2 apresentada abaixo,

$$\dot{Q}_{GV} = (h_s - h_e).m_{ag} \quad (2.2)$$

onde h_s é a entalpia do fluido na saída do Gerador de Vapor (GV), h_e é a entalpia do fluido na entrada do GV.

O valor da entalpia do fluido depende da temperatura e pressão do meio, ou seja, na entrada e na saída do GV, que em condições estáveis ou em regime permanente pode-se assumir que são constantes. Como os processos de medição de pressão e temperatura possuem boa precisão e a pressão e temperatura são constantes, pode-se assumir que a medida da entalpia também é constante e conhecida com boa precisão.

Assim, a propagação de erro no cálculo da carga térmica do GV depende da medição da vazão de água de alimentação. Entretanto, o processo de medição de vazão possui uma incerteza significativa, sendo que, a precisão, dependendo do método de

medida, pode variar de 0,5% a 2% e em alguns casos pode chegar a 5%. de erro (ANDRADE *et al.*, 2002).

Em valores típicos para uma usina nuclear do tipo PWR com 4 loops, (1350MWe) a incerteza na medida da carga térmica considerando os 4 GVs é cerca de 47 MWtérmicos (~15 MWelétricos), suponho uma incerteza de 1,5% na medida de vazão de água de alimentação.

Convém ressaltar que as cargas térmicas relativas ao circuito primário se mantêm dentro de certos valores com muito pequenas variações, tendo pouca influência na propagação de erro na medida da potência térmica do reator.

Por conseguinte, conclui-se que devido à incerteza de medição de vazão de água de alimentação, há uma incerteza significativa diretamente proporcional a vazão de água de alimentação, que se propaga no cálculo da potência térmica do reator (P_T).

Através da realização de balanços de massa e energia utilizando diversas equações redundantes e medições é possível obter valores mais exatos da vazão de água de alimentação, com isso melhorando a precisão na determinação da Potência Térmica do Reator.

Entretanto, o fechamento dos balanços de massa e energia é difícil e nem sempre é possível, devido a diversas incertezas no processo de modelagem e medição. Para contornar esse problema, pode-se utilizar a técnica de Reconciliação de Dados (RD) para realizar esse fechamento.

A técnica de Reconciliação de Dados clássica consiste basicamente em minimizar o erro quadrático, sujeito às restrições do processo. Uma vez aplicada essa técnica, os valores reconciliados ou valores corrigidos são mais exatos e com uma variância determinada. Dessa forma, após a reconciliação de dados, os valores corrigidos são aplicados ao cálculo do balanço de massa.

Convém observar na equação (2.2) que o valor da vazão de água de alimentação foi corrigido a fim de satisfazer as restrições do processo e, assim, o mesmo corresponde a um valor mais exato e com uma variância pré-determinada. Como os valores das entalpias de entrada e de saída de cada GV são considerados constantes e

medidos com muita exatidão, consideramos o valor da carga térmica do GV como o produto de um valor constante por uma variável aleatória, m_{ag} , que é mais exata e conseqüentemente o valor calculado da carga térmica também será mais exato.

Assim, ao se utilizar o valor da vazão de água de alimentação reconciliado na equação (2.2), o resultado será uma melhor determinação do cálculo da potência térmica do reator (P_t).

Para efeito de simplificação, neste trabalho não será considerado o balanço de energia, a menos quando explicitamente indicado, e como as cargas térmicas relativas ao circuito primário não tem variação considerável no cálculo da potência térmica do reator, as mesmas serão consideradas constantes e a potência térmica do reator (P_t) será equivalente à carga térmica do GV a menos de uma constante. Dessa forma, quanto mais exato o cálculo da carga térmica do GV, mais exato será o cálculo da potência térmica do reator.

Na próxima seção será apresentado um exemplo simplificado, mas realista, do balanço de massa do circuito secundário de uma usina nuclear típica do tipo PWR, o qual servirá de base para a aplicação do método desenvolvido nesse trabalho que é a Reconciliação Robusta de Dados e Seleção de Modelo Simultânea (RDSMS) no cálculo da potência térmica do reator.

2.3 – Cálculo do Balanço de Massa

A figura 2.2 abaixo mostra o diagrama simplificado do circuito secundário de uma planta tipo PWR com apenas 2 GVs baseado na norma VDI 2048 (2000) e que indica os pontos de medição e as possíveis perdas na tubulação.

As variáveis medidas são a vazão de vapor do gerador de vapor 1 (m_{GV1}), a vazão de vapor do gerador de vapor 2 (m_{GV2}), a vazão de água de alimentação 1 (m_{ag1}), a vazão de água de alimentação 2 (m_{ag2}), a vazão total de vapor (m_v), a vazão de condensado (m_c), a vazão das extrações A7, A6 e A5 (m_{A7} , m_{A6} , m_{A5}), a vazão de retorno de condensado de alta pressão (m_{HPC}) e a vazão de vapor da turbina de alta para a turbina de baixa (m_T).

A vazão de vapor no ponto de entrada da turbina de alta pode ser determinada de três diferentes formas (VDI-2048, 2000):

$$M1 = m_{GV1} + m_{GV2} - 0,2.m_v \quad (2.3)$$

$$M2 = m_{ag1} + m_{ag2} - 0,6.m_v \quad (2.4)$$

$$M3 = m_c + m_{A7} + m_{A6} + m_{A5} + 0,4.m_v \quad (2.5)$$

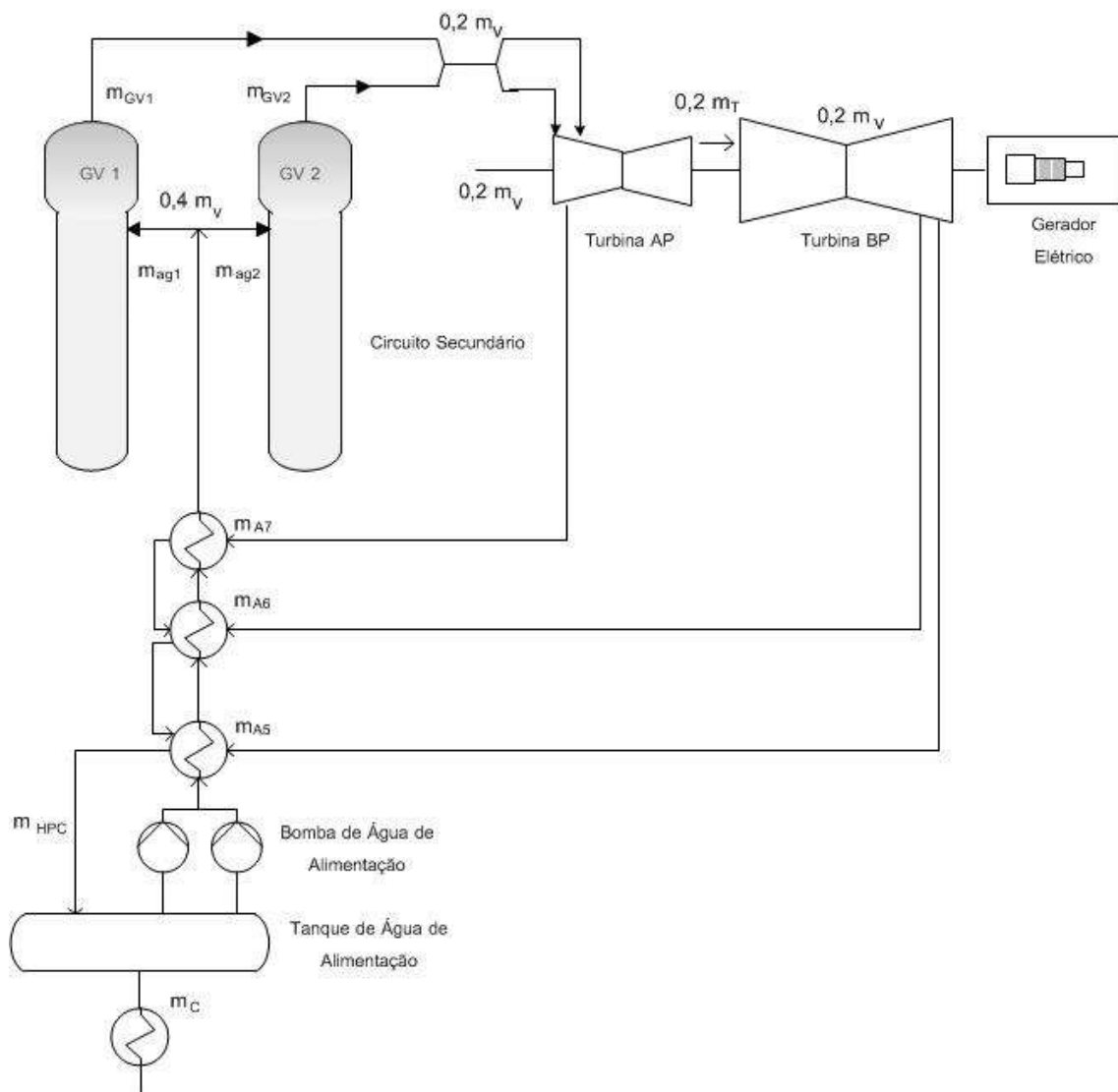


Figura 2.2: Diagrama Simplificado do Circuito Secundário de uma Usina Nuclear tipo PWR

A equação (2.3) indica a vazão de vapor que chega à turbina de alta pressão descontadas as perdas localizadas nas linhas de vapor (20%). A equação (2.4) indica a vazão de vapor que deve ser equivalente à vazão de água de alimentação total descontada as perdas nos GVs e linhas de vapor (40%+20%). A equação (2.5) indica a vazão de vapor que deve ser equivalente à vazão de condensado considerando as perdas pelas extrações e as perdas na turbina.

Uma vez determinadas as equações relativas à vazão de vapor na entrada da turbina de alta é necessário estabelecer uma relação entre as medidas que irá fornecer o balanço de massa do sistema, ou equações auxiliares ou restrições do processo a fim de serem utilizadas no processo de reconciliação de dados, as quais estão indicadas abaixo:

$$M1 - M2 = 0 \quad (2.6)$$

$$M2 - M3 = 0 \quad (2.7)$$

$$m_{A7} + m_{A6} + m_{A5} - m_{HPC} = 0 \quad (2.8)$$

As equações (2.6) e (2.7) representam a vazão de vapor na entrada da turbina de alta, dessa forma os seus valores devem ser iguais. A equação (2.8) indica que a vazão que entra no tanque de água de alimentação pela linha de retorno deve ser igual à soma da vazão de cada extração.

Uma vez determinadas as restrições do processo pode-se então exemplificar numericamente o balanço de massa antes e depois da reconciliação, utilizando como base o exemplo da norma VDI-2048 (2000). Ao aplicar os valores típicos de cada variável, essas equações apresentam resultados contraditórios, apesar de equivalentes e nesse caso o método de reconciliação de dados precisa ser aplicado VDI-2048 (2000).

Abaixo estão exemplificados os resultados utilizando os valores típicos indicados na norma VDI-2048 (2000), os valores de M1, M2 e M3 são contraditórios:

$$M1 = 91,804 \pm 1,232 \quad (2.9a)$$

$$M2 = 88,579 \pm 0,895 \quad (2.9b)$$

$$M3 = 88,687 \pm 0,875 \quad (2.9c)$$

Esses valores são contraditórios, visto que numericamente são diferentes ($M1 \neq M2 = M3$), enquanto deveriam ser iguais a menos de um intervalo de confiança, pois representam a mesma medida.

Pode-se notar que ao utilizar os valores indicados na equação (2.9), o balanço de massa indicado nas equações (2.6) e (2.7) não fecha, ou seja, as equações não se igualam.

Após a aplicação do método de reconciliação de dados, os valores obtidos conforme indicado na norma VDI-2048 (2000) correspondem ao valor médio da variável somado a um valor que indica o intervalo de confiança e estes estão indicados abaixo.

$$M1 = 88,714 \pm 0,613 \quad (2.10a)$$

$$M2 = 88,714 \pm 0,613 \quad (2.10b)$$

$$M3 = 88,714 \pm 0,613 \quad (2.10c)$$

Os valores acima estão livres de contradição, visto que os balanços de massa agora satisfazem as equações (2.6) à (2.8). Dessa forma temos maior confiabilidade na medida, podendo fazer o cálculo da potência do reator com dados mais precisos.

Para efetuar o cálculo da potência térmica do reator utilizam-se então os valores da vazão de água de alimentação reconciliados e aplicam-se na equação (2.2).

O exemplo numérico apresentado permite visualizar a possibilidade real de melhora no cálculo da potência térmica do reator utilizando a técnica de reconciliação de dados, mas não apresenta o método em si. Portanto, no próximo capítulo será apresentado o método clássico de reconciliação de dados, que servirá de base para o desenvolvimento das novas técnicas apresentadas neste trabalho.

CAPÍTULO 3:

Reconciliação de Dados e Identificação de Erros Grosseiros

3.1 - Introdução

Neste capítulo será apresentada a formulação geral do problema da Reconciliação de Dados e Identificação de Erros Grosseiros, abordando ainda a metodologia indicada na norma VDI-2048 (2000), aqui denominada como o método Clássico de Reconciliação de Dados (RDC). Ao final do capítulo serão feitas considerações sobre outros métodos para Reconciliação de Dados e Identificação de Erros Grosseiros, que serviram de base para o desenvolvimento das novas técnicas desenvolvidas neste trabalho.

A formulação geral do problema da Reconciliação de Dados é definida como um problema de minimização com restrições da seguinte forma,

$$\begin{aligned} \min_{x,u,p} \mathfrak{J}(x^M, x) \\ h(x, u, p) = 0 \quad x^{LI} \leq x \leq x^{LS} \\ g(x, u, p) \leq 0 \quad u^{LI} \leq u \leq u^{LS} \\ p^{LI} \leq p \leq p^{LS}, \end{aligned} \tag{3.1}$$

onde , \mathfrak{J} é a função objetivo, a qual depende da diferença entre as variáveis medidas x^M e os valores estimados das variáveis x , p é o conjunto de parâmetros do problema, u é o conjunto das variáveis não medidas, h é o conjunto de restrições de igualdades, g é o conjunto de restrições de desigualdade e LI e LS indicam os limites inferiores e superiores das variáveis.

Diversos métodos de RD assumem que a distribuição de probabilidade do erro é do tipo Normal com média zero e dispersão conhecida. Assim, a função objetivo adquire a seguinte forma:

$$\mathfrak{J} = (\mathbf{x}^M - \mathbf{x})^T \cdot \mathbf{V}^{-1} \cdot (\mathbf{x}^M - \mathbf{x}), \quad (3.2)$$

onde \mathbf{V} é a matriz de covariância. Normalmente a matriz de covariância não é conhecida e deve ser estimada a partir de dados do processo. Considerando o princípio de máxima verossimilhança para a distribuição Normal, um problema equivalente é minimizar a função objetivo \mathfrak{J} a seguir, sujeito às restrições dadas pela equação (3.1).

$$\mathfrak{J} = (\mathbf{x}^M - \boldsymbol{\mu})^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x}^M - \boldsymbol{\mu}) \quad (3.3)$$

onde, $\boldsymbol{\Sigma}$ é a verdadeira matriz de covariância e $\boldsymbol{\mu}$ é a verdadeira média relativa às medidas \mathbf{x}^M .

Ao comparar as equações (3.2) e (3.3) pode-se observar que o problema de RD é equivalente a determinar a variável x , que pela equação (3.3) é a média dos valores medidos, a qual por sua vez depende da estimativa da matriz de covariância do processo $\boldsymbol{\Sigma}$ (FELDMAN, 2007).

Nota-se que quando os dados estão corrompidos por erros grosseiros, um desvio é introduzido na estimativa da matriz de covariância $\boldsymbol{\Sigma}$ e no valor médio $\boldsymbol{\mu}$, o que afeta desfavoravelmente o processo de reconciliação de dados. Portanto, um passo importante para evitar erros nos dados reconciliados, os quais dependem da estimativa da matriz de covariância ($\boldsymbol{\Sigma}$) é a prévia remoção do erro grosseiro ao aplicar a técnica de reconciliação de dados.

3.2 – Reconciliação de Dados Clássica (RDC)

Nesta seção será apresentado o método de reconciliação de dados baseado na norma alemã VDI-2048 (2000).

A técnica padrão de RD baseia-se na determinação de vetor de correção para as medidas de processo a partir da minimização do erro quadrático, o qual está sujeito às

restrições das equações do processo. Nesse caso, pressupõe-se que as variáveis estão contaminadas com um ruído que possui uma distribuição de probabilidade conhecida (p. ex., distribuição Normal). A Identificação de Erros Grosseiros (IEG) utiliza a matriz de covariância do erro e testes de chi-quadrado para determinar um patamar que identifique medições errôneas, as quais devem ser eliminadas antes de se aplicar a RD.

Deve-se observar que a primeira etapa do processo de reconciliação de dados é a aquisição dos valores medidos e esses valores são considerados contaminados com um ruído aleatório e eventualmente com erros grosseiros, os quais por sua vez podem ter uma parte do EG com magnitude ou influência conhecida e outra parte com magnitude ou influência desconhecida (VDI-2048, 2000).

Os erros grosseiros cuja magnitude é conhecida podem ter sua influência descontada ou abatida previamente à aplicação da RD.

Entretanto, ao EG cuja influência é desconhecida não se pode aplicar uma compensação antes da aplicação da RD. Assim, os EGs não determinados devem ser considerados como Erros Aleatórios (EA) durante o processo de RD. Dessa forma os erros aleatórios e a parte desconhecida dos erros sistemáticos pelo menos dentro de certos limites são considerados com uma Variável Aleatória com função de distribuição contínua (VDI-2048, 2000)

Cada variável do processo x_i é considerada uma variável aleatória e, portanto pode-se considerar o conjunto das n variáveis medidas como um vetor \mathbf{x} de variáveis aleatórias de dimensão n como indicada abaixo.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (3.4)$$

Observando a equação (3.4) podemos representar graficamente o problema da RD conforme indicado na equação (3.1) da seguinte forma:

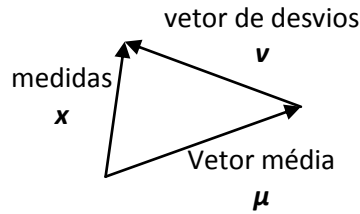


Figura 3.1: Representação gráfica da técnica de RD clássica.

O objetivo da técnica de RD é minimizar o vetor de desvios (v) sujeito a restrição das equações do processo.

Para uma variável aleatória x_i temos que a variância ($\sigma^2_{x_i}$) é a medida da incerteza associada a essa VA. Dado que x é um vetor de variáveis aleatórias, a medida da incerteza do conjunto total das medidas é dada pela matriz de covariância (Σ_x). Nota-se que os valores verdadeiros das variâncias e covariâncias não são conhecidos e precisam ser estimados.

3.2.1 – Estimação da matriz de covariância verdadeira

A estimativa da matriz de covariância verdadeira Σ_x , relativa aos valores medidos de x apresentado na equação (3.4), é constituída pelos valores estimados das variâncias e covariâncias, chamadas amostrais,

$$S_X = \begin{bmatrix} S^2_{x_{1,1}} & \cdots & S^2_{x_{1,n}} \\ \vdots & \ddots & \vdots \\ S^2_{x_{n,1}} & \cdots & S^2_{x_{n,n}} \end{bmatrix}, \quad (3.5)$$

sendo que a matriz de covariância estimada S_X deve ser uma matriz simétrica e positiva definida.

Quando as medidas de cada componente do processo são randômicas de um ciclo a outro, pode-se estimar a variância e a covariância de cada variável por meio do estimador de máxima verossimilhança.

Dado a existência de m amostras das variáveis x_i e x_k , os valores esperados estimados destas variáveis são dados por,

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^m x_{ij} \quad (3.6a)$$

$$\bar{x}_k = \frac{1}{m} \sum_{j=1}^m x_{kj} \quad (3.6b)$$

A estimativa da variância da variável x_i indicado por $S_{x_i,i}^2$ também obtida por meio do estimador de máxima verossimilhança utilizando diretamente as medidas do processo é dada pela expressão abaixo.

$$S_{x_i,i}^2 = \left(\frac{1}{m} \right) \left(\frac{1}{m-1} \right) \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2 \quad (3.7)$$

A estimativa da covariância entre duas variáveis x_i e x_k , indicado por $S_{x_i,k}^2$ pode ser obtida pela expressão apresentada na equação (3.8)

$$S_{x_i,k}^2 = \frac{1}{m} \left(\frac{1}{m-1} \right) \sum_{j=1}^m (x_{ij} - \bar{x}_i)(x_{ik} - \bar{x}_k) \quad (3.8)$$

Em alguns casos não é possível realizar a estimativa da covariância entre variáveis devido à presença de influências que não podem ser medidas, assim as covariâncias devem ser estimadas utilizando duas propriedades. Primeiro, caso duas variáveis do vetor X não sejam correlacionadas, ou seja, não existe dependência estocástica entre as variáveis x_i e x_k os valores da covariância entre x_i e x_k (S_{xik}) são nulos.

Segundo, caso alguma variável possua correlação com outra, pode-se estimar empiricamente o valor da correlação entre duas variáveis por meio do coeficiente de correlação que é definido como:

$$\rho(x_i, x_k) = \frac{Cov(x_i, x_k)}{\sqrt{Var(x_i) \cdot Var(x_k)}} \quad (3.9)$$

O coeficiente empírico pode ser obtido a partir das características das variáveis e do processo considerando a influência observada ou calculada de uma variável em outra e pode envolver a experiência e conhecimento de pessoas experientes no processo para determiná-lo. Dessa forma, a covariância entre as variáveis x_i e x_k pode ser calculada utilizando o coeficiente de correlação empírico ($r_{x_i,k}$) e o valor do desvio padrão de cada variável, como indicado pela expressão abaixo (VDI-2048, 2000):

$$S_{x_i,k} = r_{x_i,k} \cdot S_{x_i} \cdot S_{x_k} \quad (3.10)$$

3.2.2 – Intervalo de Confiança

A partir da premissa que os dados do processo estão contaminados com um ruído que possui uma distribuição Normal com média zero e variância conhecida, pode-se aplicar o conceito de intervalo de confiança para o conjunto de dados. Supondo que

desejamos estabelecer um intervalo de confiança para o valor verdadeiro μ da medida x_i com uma probabilidade p , obtemos a seguinte expressão:

$$P(x_i - \lambda_p \cdot \sigma_{xi} \leq \mu \leq x_i + \lambda_p \cdot \sigma_{xi}) = p \quad (3.11)$$

A expressão acima indica que o valor verdadeiro μ se situa com uma probabilidade p dentro do intervalo

$$x_i - \lambda_p \cdot \sigma_{xi} \leq \mu \leq x_i + \lambda_p \cdot \sigma_{xi} \quad (3.12)$$

Usualmente utiliza-se o intervalo de confiança com 95% de probabilidade, que é bem aceito em aplicações industriais. A partir de cálculos estatísticos o intervalo de confiança para a probabilidade de 95% é dado por

$$x_i - 1,96 \cdot \sigma_{xi} \leq \mu \leq x_i + 1,96 \cdot \sigma_{xi} \quad (3.13)$$

Na prática pode-se utilizar o intervalo de confiança ao invés da medida da incerteza dada pela variância da medida. A partir do intervalo de confiança a variância pode então ser estimada como a seguir (VDI-2048, 2000),

$$S_{xi}^2 = \left(\frac{V_{xi}}{1,96} \right)^2 \quad (3.14)$$

onde V_{xi} é o valor do intervalo de confiança.

3.2.3 – Aplicação da Reconciliação de Dados Clássica

O problema da RD foi formulado de acordo com a equação (3.1) e para efeito de demonstração o método será apresentado utilizando apenas a restrição de igualdade, conforme formulado na norma VDI-2048 (2000).

Seja o vetor de variáveis aleatórias \mathbf{X}^T e \mathbf{v} o vetor de correção ou erro para medida de \mathbf{X}^T . Sabendo que

$$h(x) = \begin{bmatrix} h_1(x) \\ \cdot \\ \cdot \\ \cdot \\ h_r(x) \end{bmatrix}, \quad (3.15)$$

onde $h(x)$ é o vetor das r equações auxiliares que representam os balanços de massa e energia, pode-se formular o problema da RD como:

$$\begin{aligned} \min \quad & \mathbf{v} \cdot \mathbf{S}_x^{-1} \mathbf{v} \\ s/ & \\ h(x + \mathbf{v}) = & \mathbf{0} \end{aligned} \quad (3.16)$$

Os valores medidos não satisfazem as condições auxiliares. Entretanto os valores corrigidos $\bar{\mathbf{x}} = \mathbf{x} + \mathbf{v}$ satisfazem.

Linearizando-se $h(x)$ temos que o problema de minimização assume a seguinte forma:

$$\begin{aligned} \min \quad & \mathbf{v} \cdot \mathbf{S}_x^{-1} \mathbf{v} \\ s/ & \\ h(x) + \frac{\partial h(x)}{\partial x} \cdot \mathbf{v} = & \mathbf{0} \end{aligned} \quad (3.17)$$

Aplicando a técnica de multiplicadores de Lagrange, podemos obter após diversas manipulações algébricas uma fórmula para o vetor de correção v e para a matriz de covariância estimada do erro (S_v),

$$v = -S_x \cdot H^T \cdot (H \cdot S_x \cdot H^T)^{-1} \cdot h(x) \quad (3.18)$$

$$S_v = S_x \cdot H^T \cdot (H \cdot S_x \cdot H^T)^{-1} \cdot H \cdot S_x \quad (3.19)$$

onde ,

$$H = \frac{\partial h(x)}{\partial x} \quad (3.20)$$

e

$$\frac{\partial h(x)}{\partial x} = H = \begin{bmatrix} \left[\frac{\partial h_1(x)}{\partial x_1} \right] & \dots & \left[\frac{\partial h_1(x)}{\partial x_n} \right] \\ \cdot & & \cdot \\ \cdot & \dots & \cdot \\ \cdot & & \cdot \\ \left[\frac{\partial h_r(x)}{\partial x_1} \right] & \dots & \left[\frac{\partial h_r(x)}{\partial x_n} \right] \end{bmatrix} \quad (3.21)$$

3.2.4 – Identificação de Erros Grosseiros

Após a obtenção do vetor de correção v e da matriz de covariância do erro S_v , conforme as equações (3.18) e (3.19) respectivamente, é possível determinar se o vetor de correção está dentro do intervalo de confiança desejado com a probabilidade desejada, por exemplo, com a valor de 95%.

Observa-se que o vetor de correção (v) elevado ao quadrado é uma variável aleatória com distribuição chi-quadrado, a qual depende da probabilidade desejada e do

número de graus de liberdade (número de restrições) dado pelo valor r . Dessa forma, pode-se aplicar o teste de chi-quadrado utilizando a expressão abaixo (VDI-2048, 2000).

$$(\mathbf{v})^2 \leq \chi_{r,95\%}^2 \quad (3.22)$$

Se a medida não passar no teste de chi-quadrado, indica então a presença de um erro significativo nessa componente, cujas contradições nas equações do processo estão acima do limite especificado, ou seja, com desvio muito grande (Erros Grosseiros).

Utilizando os elementos da diagonal da matriz de covariância do erro (S_v) e o tamanho do intervalo de confiança (v_i), podem-se determinar quais medidas apresentam erros significativos como indicado na expressão abaixo.

$$\left| \frac{v_i}{\sqrt{s_{v,ii}^2}} \right| \leq 1,96 \quad (3.23)$$

Os elementos que indicaram erros significativos precisam ser analisados individualmente ou em conjunto com sua variável correlacionada para se identificar a causa do desvio apresentado.

Uma vez identificado o elemento e a causa do desvio significativo, a matriz de covariância deve ser novamente calculada e o processo de reconciliação recalculado.

3.2.5 – Cálculo do Vetor de Correção e da Matriz de Covariância dos Valores Corrigidos

Após o cálculo do vetor de correção (\mathbf{v}) e da matriz de covariância do erro (S_v), devem-se calcular os valores corrigidos ou reconciliados, que são dados por

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{v} \quad , \quad (3.24)$$

onde \mathbf{x} é o vetor de medidas, \mathbf{v} o vetor de correção e $\tilde{\mathbf{x}}$ o vetor de medidas corrigido ou reconciliado.

Para determinar o intervalo de confiança da variável corrigida é necessário estimar a matriz de covariância dos valores corrigidos $\Sigma_{\tilde{\mathbf{x}}}$, que pode ser obtida a partir dos valores da matriz estimada de covariância do erro S_v e da matriz estimada de covariância da medida S_x , como indicado a seguir:

$$S_{\tilde{\mathbf{x}}} = S_x + S_v \quad (3.25)$$

Dessa forma, os intervalos de confiança dos valores corrigidos podem ser calculados utilizando os valores da diagonal da matriz $\Sigma_{\tilde{\mathbf{x}}}$, como indicado na seção 3.2.2. Deve-se notar que o intervalo de confiança dos valores corrigidos é menor do que o dos valores medidos.

3.3 – Fundamentos e Considerações sobre outros métodos de Reconciliação de Dados e Identificação de Erros Grosseiros

O método de reconciliação de dados clássico, no nosso caso, tem como principal objetivo a eliminação de contradições nos balanços de massa e energia do processo levando-se em conta que os valores a serem corrigidos precisam ter equações auxiliares redundantes.

Entretanto, o método pressupõe que as medidas são contaminadas com ruído com uma distribuição Normal com média zero e variância conhecida. Na prática essa premissa nem sempre é válida e a determinação de uma função de distribuição de probabilidade nem sempre é possível.

Ademais, a presença de erros grosseiros não explicitamente identificáveis podem alterar os valores médios e a estimativa da matriz de covariância do erro, dessa forma alterando desfavoravelmente a reconciliação de dados. Outro fator é a natureza sequencial do método clássico de Reconciliação de Dados, onde várias iterações para

remoção de erros grosseiros devem ser feitas para que se chegue a uma condição livre de erros e que a reconciliação de dados possa ser aplicada.

Com o intuito de evitar o processo iterativo utilizado para a Identificação de Erros Grosseiros (IEG) e o conseqüente desvio na estimação dos dados da planta, TJOA e BIEGLER (1991) propuseram um método para RD e IEG, onde a função objetivo incorpora o princípio de máxima verossimilhança relativo à função de Distribuição Gaussiana Contaminada, na qual são levadas em conta as contribuições devidas a erros aleatórios e erros sistemáticos.

O método consiste em minimizar uma função objetivo bivariada considerando-se um peso para cada medida, que é ajustado automaticamente em função do resíduo. Na presença de Erros Grosseiros, esse peso é menor, o que fornece uma solução sem desvio e ao mesmo tempo dá da maior robustez ao método. A solução do problema de Reconciliação de Dados é determinada por meio de um método de Programação Quadrática Sucessiva (SQP) adaptado à estrutura da função objetivo (TJOA e BIEGLER, 1991). Erros Grosseiros são identificados por meio de testes estatísticos baseados na estrutura da função objetivo

No trabalho de TJOA e BIEGLER (1991) propõe-se o uso de uma função distribuição gaussiana bivariada como indicada abaixo,

$$f = \frac{1}{\sqrt{2\pi}\sigma} \left[(1-p) \exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right) + \frac{p}{b} \exp\left(\frac{-\varepsilon^2}{2b^2\sigma^2}\right) \right] , \quad (3.26)$$

onde ε é o erro da medida, p é a probabilidade da ocorrência de erro grosseiros, b é a razão entre o desvio padrão da distribuição do erro grosseiro e o desvio padrão do erro aleatório, $b^2\sigma^2$ é a variância do erro grosseiro.

Seguindo o princípio da Máxima Verossimilhança, o problema da Reconciliação de Dados adquire a seguinte forma:

$$\begin{aligned} \min & - \sum_{\mu=1}^r \ln \left[(1-p) \exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right) + \frac{p}{b} \exp\left(\frac{-\varepsilon^2}{2b^2\sigma^2}\right) \right] \\ \text{s.t.} & \\ h(\tilde{x}, u) = 0 & \quad x_L \leq \tilde{x} \leq x_U \\ & \quad u_L \leq u \leq u_U \end{aligned} \tag{3.27}$$

A vantagem desta estratégia é que a influência do erro grosseiro é considerada no processo de reconciliação de dados e a identificação de erros grosseiros pode ser realizada de forma simultânea. Um método híbrido de Programação Quadrática Sucessiva (SQP) foi desenvolvido, cujo desempenho é mais rápido do que outros métodos de Programação Quadrática Sucessiva.

Apesar da eficiência do método híbrido referenciado acima, o problema de reconciliação de dados é não convexo e não linear e não há garantia de que haverá convergência para uma solução que seja um ótimo global (TJOA e BIEGLER, 1991; ARORA e BIEGLER, 2001).

Alguns métodos para a Reconciliação de Dados e Identificação de Erros Grosseiros possuem uma natureza combinatória. Um deles é baseado na classificação das medidas de processo em dois conjuntos: um conjunto de medidas com falha e outro conjunto de medidas sem falhas (ARORA e BIEGLER, 2001). Cada conjunto de medidas com Falha e sem Falha corresponde a um modelo estatístico e se mais de um modelo pode ser ajustado aos dados de medida, um processo de seleção de modelo será preciso para identificar o modelo correto. O processo de seleção de modelo permite de forma sistemática selecionar as medidas consideradas com falha e para a solução do problema pode-se utilizar a técnica de “Branch and Bound” ou de Programação Mista Inteira ou Programação Mista Não Linear.

YAMAMURA *et al.* (1988) propuseram um método baseado no Critério de Informação de Akaike (AIC), onde a solução do problema de identificação de erros grosseiros é resolvida minimizando-se o Critério de Informação de Akaike (AIC) e cujo modelo selecionado também satisfaz as restrições dose balanços de massa e energia. O problema foi formulado como uma estratégia de Programação Quadrática e utiliza a técnica de “Branch and Bound”.

O Critério de Informação de Akaike (AIC) é uma estimativa da distância entre o modelo de probabilidade verdadeiro ($g(x)$) e o modelo de probabilidade em consideração ($f(x,\theta)$), onde $g(x)$ é função densidade de probabilidade do modelo verdadeiro e $f(x,\theta)$ corresponde à família paramétrica de funções densidade de probabilidade, com parâmetro θ . Essa estimativa também é conhecida como critério de Kullback-Liebler de informação e possui valor positivo, a menos que os modelos sejam iguais em sua quase totalidade (AKAIKE, 1974).

O Critério de Informação de Akaike é definido como

$$AIC(\theta, k) = (-2) \cdot \text{Log}(\text{likelihood function}(\theta)) + 2k \quad , \quad (3.28)$$

onde k é o número de parâmetros ajustáveis do vetor θ . Quando existem diversas famílias de $f(x,\theta)$, ou seja, existem vários modelos, o melhor modelo ou modelo ótimo corresponde àquele que minimiza o valor de AIC (θ). Se existir apenas uma família de modelos ($f(x,\theta)$), o modelo ótimo corresponde àquele com parâmetros correspondentes ao Estimador de Mínimos Quadrados – LSE (AKAIKE, 1974). Adiante, o Critério de Informação de Akaike (AIC) terá um papel fundamental no desenvolvimento do método automático de seleção de modelo, identificação de erros grosseiros e reconciliação de dados robusta, que faz uso de um estimador redescendente robusto.

O método de reconciliação de dados e identificação de erros grosseiros proposto por YAMAMURA *et al.* (1988) considera dois tipos de classificação de erros nas medidas ou sensores: a) Sensores Normais, que correspondem a variáveis aleatórias com distribuição Normal com média zero e variância conhecida (σ^2) e b) Sensores com erros sistemáticos, que possuem desvios (μ), onde $|\mu| > \sigma^2$. Deve-se observar que se existem n instrumentos, nessas condições os erros sistemáticos são representados por uma combinação de 2^n modelos ou possibilidades. A função objetivo então é definida por,

$$\frac{1}{2} AIC(\theta) - \frac{n}{2} \ln(2\pi) \quad . \quad (3.29)$$

Assim, a formulação do problema proposta por YAMAMURA *et al.* (1988) é definida pelo problema de minimização apresentado abaixo:

$$\min_{\substack{F \in \Gamma(j) \\ y_j (j \in F) \\ x_j (j \in J)}} \frac{1}{2} \sum_{j \in J-F} x_j^2 + \frac{1}{2} \sum_{j \in F} (x_j - y_j)^2 + |F| \quad s.t. \quad (3.30)$$

$$A.x_j = b \quad j \in J \quad ,$$

onde, x_i é a medida do processo, J é o conjunto de n medidas do processo, $\Gamma(j)$ é o conjunto da potência de J , F é conjunto de instrumentos com erros sistemáticos e $|F|$ é o número de elementos em F , y_i é a razão entre o desvio correspondente aos erros sistemáticos e a variância. YAMAMURA *et al.* (1988) propuseram resolver a equação (3.30) utilizando a técnica de “Branch and Bound” e concluíram que a mesma é eficiente. O método detectou e corrigiu os erros sistemáticos nos dados medidos, mas o esforço computacional aumenta de forma significativa quando o número de sensores ou medidas com erros sistemáticos aumenta. Esse comportamento é explicado pelo aumento exponencial de combinações para a solução do problema.

SODERSTROM *et al.* (2000) propuseram um método cuja função objetivo é similar à estratégia de minimização do AIC como apresentado por YAMAMURA *et al.* (1988), mas utilizando uma estrutura para uso de Programação Inteira Mista. Nesse método os dados do processo são amostrados e organizados em uma matriz, onde as linhas correspondem aos n valores amostrados e as colunas correspondem às h amostragem dos dados em uma janela de tempo (h). Os dados são adquiridos considerando-se uma estratégia de janela móvel com horizonte h e o problema é resolvido em intervalos regulares, o que torna a estratégia particularmente interessante e prática para implementação do método de forma “On-Line” (SODERSTROM *et al.*, 2000).

Nesse método, considera-se que cada medida pode sofrer um desvio (b) e com ajustes na função objetivo é possível estimar o desvio e os valores verdadeiros das medidas provenientes do processo. Uma variável discreta (B) foi introduzida como penalidade na função objetivo a fim de representar a presença ou não de desvio. O problema foi resolvido utilizando uma estratégia de Programação Inteira Mista, o que é

computacionalmente intensivo, mas com adaptações esse esforço computacional é reduzido (SODERSTROM *et al.*, 2000). A formulação matemática é dada por:

$$\min_{\tilde{x}, B, b} \sum_{k=1}^h \sum_{l=1}^n \frac{1}{\sigma_l} |\tilde{x}_l - (x_{m_k} - b_l)| + \sum_{l=1}^n w_l B_l \quad s.t. \quad (3.31)$$

$$A.\tilde{x} = 0 \quad .$$

O sistema é representado pela matriz A , x_m são as medidas do processo, \tilde{x} são os valores estimados, b é o desvio, B é uma variável discreta indicando a presença ou não de desvio, h é o horizonte de medidas, n é o número de variáveis, σ indica o desvio padrão das medidas.

O método apresentou um bom desempenho, mas como mencionado anteriormente, o mesmo é computacionalmente intenso, devido a quantidade de combinações que devem ser processadas para se encontrar uma solução e que aumenta exponencialmente conforme é aumentada a quantidade de variáveis. Uma vantagem é que o método pode simultaneamente estimar o estado real das medidas e indicar a presença de desvio nas mesmas. Esse processo pode ainda ser estendido a processos não lineares e com características dinâmicas (SODERSTROM *et al.*, 2000).

JOHNSTON e KRAMER (1995) propuseram uma técnica para determinar o estado mais provável das medidas da planta, denominada Retificação de Dados (MLR), a qual é baseada na estimação de estado pelo princípio da Máxima Verossimilhança. O objetivo é maximizar a função de distribuição de probabilidade do estado das variáveis da planta \tilde{x} , dado o conjunto de medidas. Essa técnica explora as semelhanças ou analogia entre o princípio da máxima verossimilhança e a regressão robusta, que é uma técnica da estatística robusta desenvolvida para limitar o efeito de erros grosseiros nos dados estimados.

O método apresentado por JOHNSTON e KRAMER (1995) mostrou-se efetivo através de exemplos, que utilizam funções de distribuição gaussiana bivariada ou multivariada, pois as mesmas possuem características semelhantes às propriedades de estimadores robustos, como a função Lorentziana (JOHNSTON e KRAMER, 1995; HUBER, 1981), indicando um bom desempenho na presença de erros grosseiros e

indicando o real estado da planta, mesmo sem informações ela ou do modelo de probabilidade.

Quando os dados medidos do processo estão corrompidos por erros grosseiros, torna-se muito difícil determinar uma função distribuição de probabilidade e utilizar um estimador derivado dessa mesma função ou mesmo justificar o seu uso (ARORA e BIEGLER, 2001).

Nesses casos, com o intuito de contornar tal limitação, pode-se recorrer ao uso de estimadores robustos, pois são largamente independentes de uma função de distribuição de probabilidade específica, produzem estimativas sem desvios, mesmo na presença de um percentual significativo de desvios na medida. No processo de estimação o estimador robusto utiliza menos peso onde há desvios maiores, protegendo a influência de outras medidas com os valores corretos.

Diversos trabalhos voltados ao uso de estimadores robustos foram publicados, por exemplo, ALBUQUERQUE e BIEGLER (1996) utilizaram um M-estimador, a função FAIR para limitar o efeito causado pela presença de erros grosseiros.

ARORA e BIEGLER (2001) e WONGRAT *et al.* (2005) aplicaram com sucesso o estimador robusto de três partes de Hampel ao problema de reconciliação de dados e identificação simultânea de erros grosseiros. OZYURT e PIKE (2004) compararam técnicas de reconciliação de dados e concluíram que técnicas que utilizam estimadores robustos apresentam resultados iguais ou superiores quando comparados aos métodos seqüenciais.

No trabalho de Prata (2009) e PRATA *et al.* (2010) o estimador robusto de Welsch foi utilizado com sucesso e em VALDETARO e SCHIRRU (2009) verificou-se que os resultados obtidos usando o estimador de três partes de Hampel foram também efetivos. Dessa forma, indicando o potencial de uso desses estimadores.

Deve-se ressaltar que neste capítulo, os principais métodos que fundamentam o método automático de Reconciliação de Dados Robusta, Identificação de Erros Grosseiros e Seleção de Modelo foram citados, mas como leitura complementar, onde um levantamento bibliográfico bem completo foi realizado, pode-se consultar o trabalho de PRATA (2009), PRATA *et al.* (2009) e PRATA *et al.* (2010).

Para apresentar uma visão ampliada do problema de estimação robusta, no próximo capítulo será apresentado um resumo sobre estatística robusta e sobre o uso de estimadores robustos.

CAPÍTULO 4:

Estatística Robusta e Estimadores Robustos

4.1 - Introdução

A maioria dos métodos de reconciliação de dados e identificação de erros grosseiros pressupõe a existência de uma função distribuição de probabilidade conhecida, usualmente a função de probabilidade Gaussiana, com média zero e variância conhecida (σ^2). Quando as medidas do processo estão contaminadas pela presença de erros grosseiros, a determinação de uma função de probabilidade se torna muito difícil, bem como, o uso de um estimador derivado dessa função (ARORA e BIEGLER, 2001).

A presença de pequenos desvios entre o modelo de probabilidade real e o modelo considerado, devido aos dados corrompidos por erros grosseiros, pode influenciar drasticamente e negativamente os valores estimados. Assim, nesse caso, as técnicas da estatística robusta devem ser utilizadas.

A estatística robusta foi desenvolvida para lidar com tais desvios do modelo de probabilidade real e seus desdobramentos e conseqüências, sendo que estimativas estáveis e confiáveis podem ser obtidas com técnicas robustas, quando desvios do modelo de probabilidade real ocorrem até um determinado ponto (RONCHETTI, 1997b).

A base da estatística robusta remonta ao ano de 1964 relacionada ao trabalho pioneiro de HUBER (1964, 1981), onde uma generalização do estimador de máxima verossimilhança foi estabelecida e os estimadores com determinadas propriedades foram denominados M-Estimadores (HAMPEL *et al.*, 1986).

HUBER (1981) propôs obter estimadores baseado na idéia de substituir o quadrado do erro por outra função com a forma,

$$\min_T \sum_{i=1}^m \rho(x_i - T) \quad , \quad (4.1)$$

onde ρ é não constante e possui uma função derivada (ψ), de forma que o estimador T satisfaça a seguinte equação,

$$\sum_{i=1}^m \psi(x_i - T) = 0 \quad , \quad (4.2)$$

onde ψ é a derivada de ρ em relação a x , x_i é o vetor de variáveis medidas. Quando a função ρ corresponde a $-\ln(f_0(x))$, onde $f_0(x)$ é a função de distribuição Normal, o M-Estimador é o Estimador de Máxima Verossimilhança - MLE (HAMPEL *et al.*, 1986).

Adiante neste capítulo será apresentada primeiramente uma visão sucinta sobre estatística robusta, em especial sobre os estimadores robustos, onde serão mostrados alguns estimadores comumente utilizados, além de outros aspectos sobre a estabilidade global dos M-Estimadores e outros parâmetros que fornecem informações qualitativas e quantitativas sobre estimadores robustos.

4.2 – Estatística Robusta e Estimadores Robustos

A análise estatística utiliza como ferramenta poderosa a regressão linear e o método de estimação dos mínimos quadrados, que é um método amplamente difundido e utilizado. Apesar disso, o método de mínimos quadrados possui falta de robustez, pois, apenas um desvio em uma das medidas pode alterar os valores estimados de forma significativa (HAMPEL *et al.*, 1986).

Dessa forma apareceu a necessidade de buscar novos estimadores capazes de tratar os dados estatísticos com algum grau de contaminação ou presença de Erros Grosseiros.

Os resultados estatísticos relativos à eficiência e sensibilidade de estimadores em relação à presença de erros grosseiros puderam ser medidos a partir de resultados propostos por Hampel (ROUSSEEUW e LEROY, 1987), os quais foram publicados em 1974 (HAMPEL, 1974) na forma de uma “função de influência”. Diversos outros trabalhos foram publicados posteriormente, trazendo resultados importantes sobre os M-Estimadores.

Um importante conceito é que a função de influência descreve o efeito de uma contaminação infinitesimal em um determinado ponto x e que permite avaliar a partir do comportamento infinitesimal o desvio assintótico causado pela contaminação nas observações (OZYURT e PIKE, 2004).

A Função de Influência (IF) de um estimador T sobre uma determinada função de distribuição F é dada por,

$$IF(x, T, F) = \lim_{t \rightarrow 0} \frac{T(F_t) - T(F)}{t} \quad . \quad (4.3)$$

Onde F_t representa a função de distribuição no tempo t .

Simplificadamente, a função de influência (IF) é a primeira derivada ordinária de uma função de distribuição (F) relativa a um ponto x em um espaço de probabilidades de dimensão infinita (HAMPEL *et al.*, 1986).

A função de Influência fornece importantes índices de desempenho em relação aos M-estimadores, que fornecem dados qualitativos e quantitativos sobre o comportamento dos mesmos.

Os índices mais significativos do ponto de vista da robustez são apresentados a seguir (HAMPEL *et al.*, 1986):

- a) *Sensibilidade a erros grosseiros* (γ), que é definido como o supremo do valor absoluto da função de influência em todo o domínio. O valor de γ mede o efeito que uma determinada contaminação tem sobre o estimador;
- b) *Sensibilidade em relação a desvio de medida* (“*local-shift sensivity*”), que mede qual o limite da influência no estimador se um valor x se afastar do seu

valor original (“*leverage point*”). O valor da sensibilidade em relação a desvio de medida (λ) é dado pela inclinação da função de influência no ponto considerado.

- c) *Ponto de rejeição* (“*rejection point*”), que está relacionado com a rejeição completa de erros grosseiros extremos. O ponto de rejeição ζ é definido como o ponto $c > 0$ tal que a função de influência é nula para todo valor absoluto de x maior do que c . Todas as observações acima desse valor serão rejeitadas completamente.

O estudo desses índices permite entender melhor os problemas relacionados aos M-estimadores e desenvolver outros estimadores que possuam comportamento específico (HAMPEL *et al.*, 1986).

A *sensibilidade a erros grosseiros* γ é dada por,

$$\gamma = \sup_x |IF(x, T, F)| \quad . \quad (4.4)$$

A definição acima pode ser vista como um limite aproximado do desvio do estimador e, a partir dessa definição, um importante passo para dar mais robustez a uma estimador é limitar o valor absoluto da função de influência (HAMPEL *et al.*, 1986).

A *sensibilidade em relação a desvio de medida* (“*local-shift sensitivity*”), é calculada como,

$$\lambda = \sup_{x \neq y} |IF(y, T, F) - IF(x, T, F)| / |y - x| \quad . \quad (4.5)$$

O ponto de rejeição ζ é calculado segundo,

$$\zeta = \inf \{c > 0; IF(x, T, F) = 0 \text{ para } |x| > c\} \quad . \quad (4.6)$$

onde ζ é o menor valor de $|x|$ tal que a Função de Influência é nula a partir de um determinado ponto.

A figura 4.1 ilustra a definição dos três pontos indicados acima.

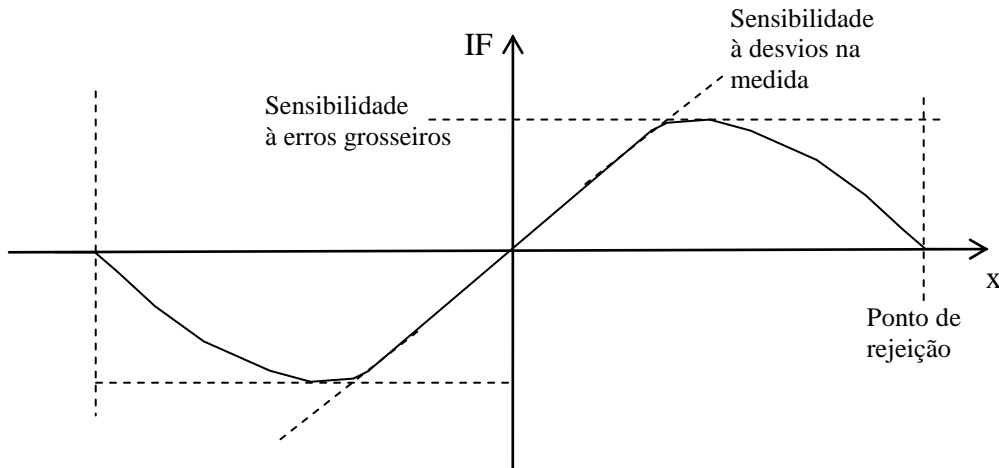


Figura 4.1: Função de Influência e os seus pontos característicos.

A função de influencia permite a definição de vários índices de desempenho de M-estimadores, mas não fornece nenhuma informação em relação à estabilidade global desses estimadores (HAMPEL *et al.*, 1986). Assim torna-se necessário complementar os índices de desempenho dos M-estimadores com a informação sobre sua estabilidade global definindo o conceito de ponto de ruptura (“*breakdown point*”), que indica o ponto em que o estimador é totalmente não confiável, ou seja, a medida da menor fração de contaminação que não influencia a estatística considerada.

Sejam m medidas na amostra de dados, k o número de medidas que estão contaminadas e T o estimador considerado.

O ponto de ruptura é definido por

$$\varepsilon(T) = \min \left\{ \frac{k}{m}, \mid bias(T) \text{ é infinito} \right\} \quad (4.7)$$

onde $bias(T)$ é dado pelo desvio assintótico do estimador T e $\frac{k}{m}$ é a razão entre as medidas contaminadas da amostra e o total de amostra, ou seja, o ponto de ruptura é um percentual de contaminação em relação às medidas e seu valor é definido como o valor

mínimo em que o grau de contaminação, ou seja, a quantidade de amostras contaminadas em relação ao total de amostras, que pode fazer com que os valores estimados se desviem do valor real da medida por um valor arbitrariamente distante e com a magnitude que se queira, perdendo assim sua característica de robustez.

Como exemplo, pode-se utilizar o estimador de mínimos quadrados, onde o valor esperado da amostra é dado pelo valor da média das m amostras. Observa-se que apenas com um erro grosseiro pode-se alterar o valor estimado tanto quanto se queira. Por conseguinte, o ponto de ruptura é dado por:

$$\varepsilon(T) = \frac{1}{m} \quad , \quad (4.8)$$

sendo que o valor de ruptura tende a zero conforme o número de medidas aumenta. Neste caso, o ponto de ruptura é de 0%. A média de uma amostra de dados é um estimador que não é limitado, pois ao se substituir apenas uma medida por um valor arbitrário, a média sofrerá um desvio tão maior quanto maior for o desvio da medida. Entretanto, a mediana é um estimador que possui um valor de ruptura de 50%, ou seja, o estimador é limitado até quando pelo menos 50% das observações forem alteradas (ROUSSEEUW e CROUX, 1993).

Outros estimadores foram propostos para contornar o problema de robustez do estimador de mínimos quadrados, como por exemplo, o estimador de mínimo valor absoluto (norma L_1). Entretanto, esse estimador também possui ponto de ruptura de 0%, pois o mesmo não protege de erros grosseiros no eixo x das abcissas (“*leverage point*”), ou seja, em uma regressão a amostra consiste em um conjunto de pontos ordenados (x_i, y_i) e o erro da medida pode se apresentar tanto no valor das ordenadas ($y_i=y+\delta_y$) quanto no eixo das abcissas ($x_i=x+\delta_x$). O estimador robusto deve apresentar a característica de proteger a estimativa na presença de erros tanto nos valores das ordenadas quanto nos valores das abcissas.

Um estimador robusto importante, muito utilizado na estatística robusta, é o estimador MAD (“Median Absolute Deviation”) denominado Desvio Absoluto da Mediana, que definido por,

$$MAD(x_i) = \text{mediana}_i \left\{ \left| x_i - \text{mediana}_i(x_j) \right| \right\}. \quad (4.9)$$

O estimador MAD possui ponto de ruptura de 50% com função de influência (IF) mais precisa sobre as diversas classes de estimadores (ROUSSEEUW e CROUX, 1993). O cálculo do MAD é de fácil implementação e costuma ser utilizado como estimador robusto do fator de escala (S_n) para normalização dos resíduos dos estimadores robustos. O cálculo do fator de escala S_n (desvio padrão – σ) utilizando MAD é dado pela equação abaixo:

$$S_n = 1,483MAD(x_i) = 1,483 \text{mediana}_i \left\{ \left| x_i - \text{mediana}_i(x_j) \right| \right\}. \quad (4.10)$$

Diversos estimadores têm sido desenvolvidos e são encontrados na literatura, como por exemplo, em HUBER (1981), BASU e PALIWAL (1989), HAMPEL (1974) e HAMPEL *et al.* (1986). Dentre os diversos estimadores, destacamos a função FAIR utilizada por ALBUQUERQUE e BIEGLER (1996) e definida como,

$$\rho(\varepsilon, c_f) = c_f^2 \left[\frac{|\varepsilon|}{c_f} \right] + \log \left(1 + \frac{|\varepsilon|}{c_f} \right), \quad (4.11)$$

onde ε é o erro da medida e c_f uma constante que deve ser ajustada para regular a eficiência e a robustez do estimador. O ajuste da constante c_f é feito com base em um compromisso entre esses dois parâmetros (ALBUQUERQUE e BIEGLER, 1996), onde c é proporcional a uma função da Eficiência Assintótica E ($c=0.21529.f(E)^{1.02}$).

Outro estimador com destaque é o estimador Normal Contaminada, que é definido como,

$$\rho(\varepsilon, p, b, \sigma) = -\ln \left[(1-p) \exp \left(\frac{-\varepsilon^2}{2\sigma^2} \right) + \frac{p}{b} \exp \left(\frac{-\varepsilon^2}{2b^2\sigma^2} \right) \right], \quad (4.12)$$

onde p é a probabilidade e $b^2\sigma^2$ é a variância relativa a contaminação pelo erro grosseiro (OZYURT e PIKE, 2004).

Uma subclasse de estimadores cuja função de influência é nula para todos os valores acima de um determinado valor c , a qual é capaz de rejeitar erros grosseiros acima desse limite, é a dos estimadores redescendente. Esses estimadores possuem as três características relacionadas à robustez mencionadas nos itens (a), (b) e (c) acima e vários estimadores do tipo redescendente já haviam sido propostos antes da formalização da função de influência detalhada em HAMPEL (1974).

Entre os estimadores redescendentes temos a função seno de Andrew, a função Biweight de Tukey e os estimadores de duas partes e de três partes de Hampel. Esse último teve um papel de destaque no estudo sobre robustez conduzido pela Universidade de Princeton (HAMPEL *et al.*, 1986).

Diversos autores utilizaram o estimador de três partes de HAMPEL com sucesso no problema de reconciliação de dados e na identificação de erros grosseiros, entre eles podemos citar ARORA e BIEGLER (2001), WONGRAT *et al.* (2005) e VALDETARO e SCHIRRU (2009, 2011). Na figura 4.2 é mostrada a representação gráfica do estimador de três partes de Hampel.

A figura 4.2 abaixo indica graficamente a FI do estimador de Hampel de três partes, o que permite fazer uma comparação qualitativa com a figura 4.1 e visualizar os três itens relacionados à robustez.

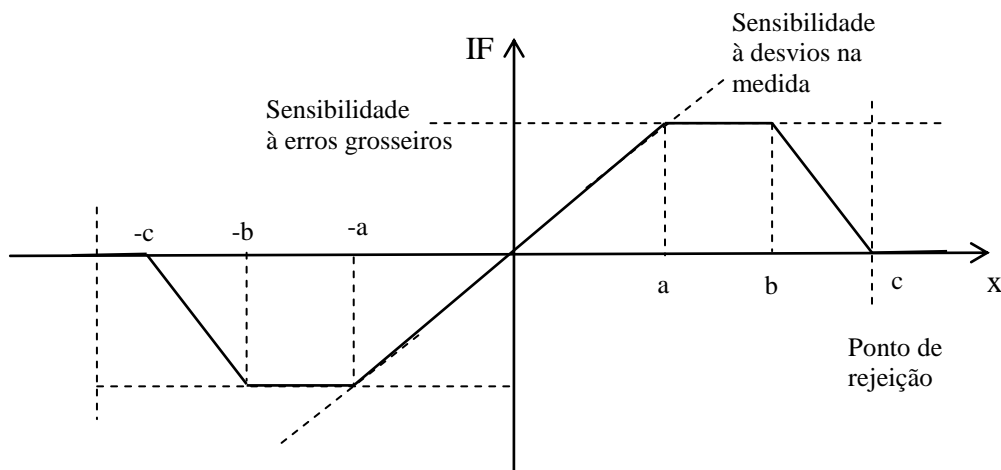


Figura 4.2: Função de Influência do Estimador de Três Partes de Hampel.

Na próxima seção será apresentado o estimador robusto de Hampel e serão feitas outras considerações em relação a esse estimador.

4.3 – Estimador Robusto Redescendente de Três Partes de Hampel

O estimador de três partes de Hampel é um estimador do tipo M (máxima verossimilhança) e tem associado sua função característica ψ , que quantifica o efeito do resíduo em relação aos dados estimados.

Quando a função característica ψ é nula a partir de um valor c , o estimador é classificado como sendo do tipo redescendente. Este tipo de estimador possui a propriedade de completa rejeição a erros Grosseiros, o que é uma característica importante que pode ser aproveitada na resolução do problema de Reconciliação de Dados.

O estimador de Hampel de três partes é apresentado na função objetivo F_H (uma função de distribuição de probabilidade) como indicado a seguir:

$$F_H \begin{cases} \frac{1}{2} \varepsilon_i^2, & 0 < |\varepsilon_i| \leq a \\ a|\varepsilon_i| - \frac{1}{2} a^2, & a < |\varepsilon_i| \leq b \\ a.b - \frac{1}{2} a^2 + (c-b) \frac{a^2}{2} \left[1 - \left(\frac{c-|\varepsilon_i|}{c-b} \right)^2 \right], & b < |\varepsilon_i| \leq c \\ a.b - \frac{1}{2} a^2 + (c-b) \frac{a^2}{2} & |\varepsilon_i| > c \end{cases} \quad (4.13)$$

onde, ε_i é o erro de cada amostra $(x_i^M - \tilde{x}_i)$, \tilde{x}_i é o valor estimado da medida e a , b e c são constantes que devem ser encontradas para que o estimador se ajuste o mais próximo possível à característica estatística dos dados. As constantes a , b , e c devem satisfazer a seguinte relação:

$$c \geq b + 2.a \quad (4.14)$$

Nota-se que no intervalo $[-a, a]$, o comportamento da f_0 do estimador é similar a função objetivo de mínimos quadrados. No intervalo $[-c, -b]$ e $[b, c]$ a probabilidade cai rapidamente e no intervalo $(-\infty, c)$ e (c, ∞) permanece praticamente constante. Essa parte constante é a que tem papel mais significativo na característica de robustez do estimador.

Esse estimador possui três constantes que devem ser ajustadas, o que aparenta ser uma característica não muito favorável. Entretanto, o ajuste dessas constantes permite sintonizar adequadamente o estimador aos dados estatísticos e também detectar erros grosseiros (WONGRAT *et al.*, 2005).

Uma vantagem significativa do estimador redescendente de três partes de Hampel é que o mesmo possui um ponto de corte explícito, c , tal que a medida é considerada um Erro Grosseiro se o resíduo for maior do que esse valor.

$$|\varepsilon_i| \geq c \quad . \quad (4.15)$$

Outros estimadores não possuem ponto de corte ou não o possuem de forma explícita, por exemplo o estimador por mínimos quadrados (LSE), a Função Fair, a Função Logística, o Estimadores de Cauchy, Huber e Welsch. Nestes casos a detecção de Erros Grosseiros é feita por critérios alternativos como mostrado a seguir.

4.4 – Diferentes Critérios para Identificação de Erros Grosseiros

Os estimadores redescendentes, aqui especificamente o estimador redescendente de três partes de Hampel, possuem um ponto de corte explícito ou “cut off point”, como apresentado na equação (4.15), onde a partir desse ponto os erros grosseiros são completamente rejeitados, ou seja, o peso à aquele desvio é nulo. Dessa forma, qualquer valor da medida além desse ponto pode ser considerado um Erro Grosseiro e sua magnitude não será significativa no cálculo dos valores estimados.

No estimador de três partes de HAMPEL esse ponto de corte é um valor explícito e bem determinado, sendo inexistente em diversos estimadores não

redescendentes. Para estes torna-se necessário o uso de outros critérios de Identificação de Erros Grosseiros, alguns dos quais estão apresentados a seguir.

No caso da aplicação do método de reconciliação de dados clássica, a IEG é feita por meio de testes estatísticos, como o teste de chi quadrado, mas por ser um teste de hipótese, o mesmo pode ter suas condições de hipótese violadas introduzindo erros de identificação (PRATA, 2009).

Quando o estimador não possui um ponto de corte explícito, outros critérios para identificação de erros grosseiros são utilizados. HAMPEL *et al.* (1986) consideram entre outros, o método identificado como X84, cuja regra de rejeição é baseada na mediana ao invés da média e desvio padrão. A regra X84 tem como princípio rejeitar qualquer desvio acima de 5,2 desvios da mediana ou da MAD, conforme a equação (4.16).

$$EG \approx |x_i^M - x_i| \geq 5,2 MAD(x_i) \quad (4.16)$$

Outro critério considerado para Identificação de Erros Grosseiros é o de Farris e Law para a Distribuição Normal Contaminada, onde o ponto de corte (c_{op}) é definido de forma equivalente como (OZYURT e PIKE, 2004),

$$\max\left\{P\left(\left(\text{medida} > c_{op}\right) \text{ e é um EG}\right) - P\left(\left(\text{medida} > c_{op}\right) \text{ e NÃO é um EG}\right)\right\}. \quad (4.17)$$

O ponto de corte do critério de Farris e Law é dado pela equação abaixo, sendo que os parâmetros p e b são ajustes relativos à contaminação por EG (PRATA, 2009):

$$EG \approx |x_i^M - x_i| \geq \sqrt{\frac{2b^2}{(b^2 - 1)} \ln\left(\frac{b(1-p)}{p}\right)} \quad (4.18)$$

Outro critério apresentado por OZYURT e PIKE (2004) é baseado em algumas características da Função de Influência do Estimador Robusto, como o ponto de máximo, ponto de mínimo, ou pontos de inflexão das derivadas primeira e segunda. Esses pontos podem ser escolhidos de forma sistemática e quanto menor o ponto de corte, a detecção de Erros Grosseiros pode ser melhorada, mas também podendo aumentar a quantidade de falsas detecções de EG e aumentar a variância do valor estimado. Convém ressaltar que o ponto de corte do critério de Farris e Law se situa na parte descendente da Função de Influência (OZYURT e PIKE, 2004).

Apesar da existência de um ponto de corte no estimador redescendente de três partes de Hampel, existem três constantes (a, b, c) que devem ser ajustadas de forma a garantir um ajuste do estimador aos dados estatísticos. Esse ajuste ou seleção de modelo de probabilidade é um ponto fundamental no estudo de estimadores robustos, visto que se o mesmo não estiver corretamente ajustado, os valores estimados poderão ser totalmente ineficazes, por exemplo, apresentando um valor de desvio (“bias”) significativo, o que pode indicar um valor de *sensibilidade a erros grosseiros* (γ) alto para a amostra.

Assim, na próxima seção será apresentada uma das formas de ajuste das constantes de um estimador robusto, a qual é baseada na minimização do índice de um critério de informação, que é uma forma eficiente e consagrada para a determinação e ajuste das constantes de estimadores. O critério utilizado é o Critério de Informação de Akaike e serão tratados alguns aspectos sobre o ajuste dos estimadores, especificamente o estimador redescendente de três partes de Hampel.

4.5 – Ajuste do Estimador Redescendente de Hampel

Os estimadores do tipo redescendentes pertencem a uma ampla família de Distribuições de Probabilidade e, dessa forma, torna-se necessário realizar o ajuste dos seus parâmetros a fim de ajustá-lo à série de amostras. No caso do estimador de Hampel é necessário ajustar as constantes a, b, e c, as quais estão indicadas na equação (4.13) de forma a ajustar o estimador ao membro correto dessa família de Distribuições de Probabilidade.

O ajuste dessas constantes pode parecer um problema dos estimadores redescendentes, mas sob outra ótica, esse ajuste do sistema que está sendo monitorado serve para melhorar a eficiência na detecção de erros grosseiros e no processo de reconciliação de dados.

Observa-se que o problema do ajuste das constantes dos estimadores robustos ainda é um problema pouco explorado e de alguma forma não foi dada toda a importância a esse assunto (RONCHETTI, 1997a). Assim, consideramos que ainda existe muito a ser explorado nessa área.

Os estimadores robustos do tipo M propostos por HUBER devem satisfazer as equações (4.1) e (4.2), mas para os estimadores redescendentes a solução de $\sum_{i=1}^n \psi(x_i - T) = 0$ não é única, pois os valores acima de c são todos nulos. Uma forma de determinar a solução desse problema é encontrar o mínimo global de $\sum_{i=1}^n \rho(x_i - T)$. Outras formas de solução podem ser selecionar soluções próximas à mediana; ou utilizar uma solução aproximada, por exemplo, usando o método de Newton-Raphson (HAMPEL *et al*, 1986).

Como os estimadores robustos pertencem a uma família de distribuições, o ajuste dos parâmetros do estimador é de fundamental importância. Na figura 4.3 pode-se verificar que ao se variar os valores das constantes, a *sensibilidade a erros grosseiros* (γ) e o ponto de rejeição são alterados. A *sensibilidade em relação a desvio de medida* (“*local-shift sensitivity*”) não se altera, devido à proporção mantida entre as constantes a , b e c , nesse exemplo (vide equação 6.6).

Dependendo do valor das constantes de ajuste a , b , e c , o valor estimado também pode mudar. Assim, se o estimador não estiver com a sintonia desses parâmetros feita adequadamente, os valores estimados serão imprecisos e ineficientes, assim como a rejeição a Erros Grosseiros.

Uma forma de realizar o ajuste adequado dos parâmetros de um estimador foi apresentada por ARORA e BIEGLER (2001) e a mesma utiliza um critério de informação para obtenção do melhor ajuste das constantes do estimador robusto de três partes de Hampel.

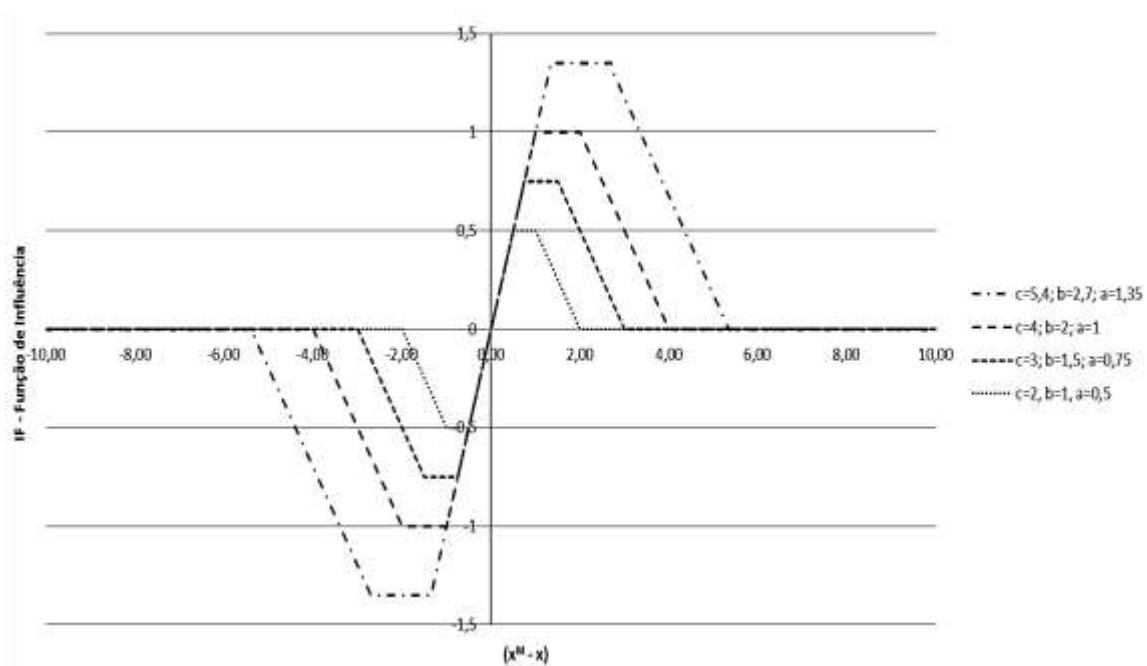


Figura 4.3: Função de Influência (IF) do Estimador de Três Partes de Hampel em função do ajuste dos parâmetros a, b e c.

ARORA e BIEGLER (2001) utilizaram o Critério de Informação de Akaike (AKAIKE, 1974) e um procedimento em dois passos e iterativo para determinar o melhor.

As constantes são escolhidas em um primeiro passo e depois de calculados os valores ótimos das constantes a, b, e c, em um segundo passo, utiliza-se essas constantes para resolver o problema da reconciliação de dados e identificação de erros grosseiros.

O objetivo do primeiro passo desse procedimento é minimizar o critério de Akaike em relação às constantes de ajuste do estimador. Inicialmente devem-se estabelecer dois conjuntos de valores das constantes a, b e c, de forma que um valor mínimo do AIC esteja entre os valores do AIC correspondentes a esses dois valores escolhidos previamente ($AIC_{abc1} < AIC < AIC_{abc2}$). Como as constantes a, b e c obedecem a uma relação de proporcionalidade (vide equação 6.6), as constantes b e a são obtidas diretamente em função do valor atribuído à constante c.

Os valores de AIC de cada extremo são obtidos por meio de testes prévios e uma vez que um valor de mínimo do AIC esteja estimado entre esses dois valores extremos

do AIC (AIC_{abc1} e AIC_{abc2}), um algoritmo de busca, no caso, o da seção áurea, é aplicado para obter as melhores constantes a, b e c que minimizem o AIC.

Após a determinação das constantes a, b e c que minimizam o AIC, elas devem ser utilizadas no estimador de três partes de Hampel e utilizadas nessa segunda parte do procedimento, na solução do problema da Reconciliação de Dados e Identificação de Erros Grosseiros.

No trabalho de WONGRAT *et al.* (2001) foi utilizada a relação obtida por YAMAMURA *et al.* (1988), onde foi pressuposto que a contaminação do erro possuía distribuição Normal com média zero e variância conhecida. Essa relação é apresentada abaixo e também foi a utilizada no trabalho de VALDETARO e SCHIRRU (2009);

$$AIC = \sum_{i=1}^{m-n_o} \left(\frac{x_i^M - \tilde{x}_i}{\sigma_i} \right)^2 + 2.n_o \quad , \quad (4.19)$$

onde x_i^M é a i-ésima variável medida, \tilde{x}_i é o valor estimado, σ_i o desvio padrão da i-ésima variável, m é o número total de medidas, n_o é o número de Erros Grosseiros detectados, O primeiro termo da equação (4.16) é o critério de ajuste, enquanto o segundo termo é uma penalidade. Após a determinação do melhor ajuste para as constantes a, b e c, pode-se utilizar a constante c como indicado na equação (4.15) para detecção de Erros Grosseiros no problema da Reconciliação de Dados.

4.6 – Considerações sobre os Estimadores Robustos e o método de ajuste das constantes do Estimador Redescendente de Hampel

Observando os conceitos sobre robustez apresentados em HAMPEL (1974), seção 4.5, os requisitos básicos para um estimador robusto são a fraca reação a pequenas perturbações, ou seja, o valor estimado não é alterado na presença de pequenas perturbações ou pequenos desvios na função de probabilidade associada à contaminação do erro e, que sejam seguros na presença de Erros Grosseiros em quantidade e com magnitude significativa, o que implica em um alto ponto de ruptura, ver equação (4.7).

Os estimadores robustos devem possuir um limite superior em relação à influência relativa de qualquer contaminação, o que corresponde a um valor mínimo da *sensibilidade a erros grosseiros* γ , ver equação (4.4). Outro ponto importante é impor uma clara separação entre o conjunto de dados medidos ou amostrados e Erros Grosseiros, o que significa um adequado e baixo ponto de rejeição ζ , ver equação (4.6), e ainda estimar eficientemente a quantidade correta.

Convém ressaltar que *sensibilidade a erros grosseiros* γ , como indicado na equação (4.4), é um índice que mede a pior influência que uma contaminação determinada pode ter sobre o valor do estimador (HAMPEL *et al.*, 1986) e sendo este finito pode ser visto como o desvio assintótico do estimador.

Na figura 4.3, pode-se observar que as características da Função de Influência do estimador de três partes de Hampel mudam ao se variar o valor das constantes a, b e c, sendo que, considerando o ajuste das constantes usando o critério de Akaike, a minimização da constante a, b e c leva a um valor mínimo à *sensibilidade a erros grosseiros* e o *ponto de rejeição*, conforme explicado acima.

Considerando-se as propriedades favoráveis descritas acima e inerentes ao Estimador de Três Partes de Hampel e que não estão presentes em uma parcela significativa de outros, será desenvolvido no próximo capítulo um método para efetuar a estratégia de ajuste das constantes do estimador de três partes de Hampel simultaneamente à reconciliação robusta de dados e a identificação de erros grosseiros.

O método proposto nesta tese é baseado na minimização direta de um índice de desempenho, que é o Critério de Informação de Akaike Robusto, e consiste em uma característica inovadora deste trabalho, a qual é aplicada ao problema de Reconciliação de Dados e à Identificação de Erros Grosseiros e apresentada adiante.

Outro ponto importante na utilização de um estimador robusto no problema da reconciliação de dados e identificação de erros grosseiros é que a mesma pode ser não linear e não convexa, e o cálculo efetuado pelo algoritmo de otimização pode levar a uma solução com um mínimo local. Dessa forma, deve-se recorrer a um algoritmo de otimização global.

Métodos de otimização baseados na teoria evolucionária como o Algoritmo Genético (GA) e o algoritmo baseado em enxame de partículas são algoritmos de otimização globais, portanto, uma alternativa promissora aos métodos tradicionais.

Baseado em diversos trabalhos utilizando algoritmos evolucionários como o Algoritmo Genético modificado proposto por WONGRAT *et al.* (2005) e da eficácia no uso do algoritmo de otimização por enxame de partículas como indicado nos trabalhos de VALDETARO e SCHIRRU (2009, 2011) e PRATA *et al.* (2009, 2010), nesta tese será utilizado no seu desenvolvimento o algoritmo de otimização por enxame de partículas padrão.

Na próxima seção, será apresentado uma visão geral do método de otimização baseada em enxame de partículas que é um algoritmo que tem se mostrado robusto quando aplicado a diversos problemas e ele será utilizado na solução do problema de RD e IEG neste trabalho, como será mostrado no capítulo 5.

Os estimadores robustos parecem uma alternativa promissora para aplicação no problema da Reconciliação de Dados e Identificação de Erros Grosseiros. Porém, a etapa de ajuste das constantes dos estimadores acrescenta uma fase a mais no problema de RD e na IEG.

4.7 – Algoritmo de Otimização por Enxame de Partículas

Diversas técnicas evolucionárias são inspiradas na evolução natural que ocorre na natureza, por exemplo, Programação Genética (PG) e Algoritmos Genéticos. Os indivíduos que pertencem a uma população trazem consigo características próprias ou codificadas, que acabarão por fazer parte da solução final do problema. Essas características são alteradas através de operações semelhantes às transformações que ocorrem nos genes de um ser vivo, por exemplo, mutação, seleção, cruzamento ou “crossover” e reprodução as quais dependem de uma função de custo ou de ajuste (“fitness”) de cada indivíduo (SHI e EBERHART, 1998).

KENNEDY e EBERHART (1995) propuseram um algoritmo diferente baseado no comportamento social de um grupo ao competir por um determinado recurso. Esse

algoritmo foi baseado no comportamento de um bando de pássaros ao disputar alimento e como já colocado é denominado Algoritmo de Otimização por Enxame de Partículas ou PSO, que é uma abreviatura do termo em inglês (“Particle Swarm Optimization”). No algoritmo PSO não ocorrem manipulações genéticas. Ao invés disso, as partículas se movimentam numa estratégia que mistura cooperação e competição entre elas (SHI e EBERHART, 1998).

A cada passo, as partículas ou indivíduos ajustam sua experiência de vôo (posição, direção e velocidade) baseadas na sua própria experiência ou nas informações da sua melhor função de custo e nas informações da melhor função de custo de todo o grupo ou na melhor informação global. Cada elemento do grupo representa uma partícula e cada uma delas é uma potencial solução para o problema.

Cada partícula é tratada como um ponto em um espaço n-dimensional e a i-ésima partícula é representada por:

$$X_i = [x_{i1}, x_{i2}, \dots, x_{in}] \quad (4.20)$$

A equação abaixo representa a taxa de variação da posição (velocidade) para a i-ésima partícula.

$$V_i = [v_{i1}, v_{i2}, \dots, v_{in}] \quad (4.21)$$

A melhor posição individual (P_i) ou a que possui a melhor função de custo para a partícula considerada está apresentada na equação (4.22) e a melhor posição global (P_g) ou a que possui o melhor ajuste (“fitness”) entre todas as partículas está representada na equação (4.23).

$$P_i = [p_{i1}, p_{i2}, \dots, p_{in}] \quad (4.22)$$

$$P_g = [x_{g1}, x_{g2}, \dots, x_{gn}] \quad (4.23)$$

A cada iteração (k) a posição e a velocidade são atualizadas conforme as equações abaixo:

$$V_i(k+1) = w.V_i(k) + c_1.rand_1.[P_i(k) - X_i(k)] + c_2.rand_2.[P_g(k) - X_i(k)] \quad (4.24)$$

$$X_i(k+1) = X_i(k) + V_i(k+1) \quad , \quad (4.25)$$

onde c_1 e c_2 são duas constantes positivas, $rand_1$ e $rand_2$ são dois números aleatórios no intervalo $[0, 1]$ e w é o fator de inércia.

No lado direito da equação (4.24) o segundo termo representa a parte “cognitiva” em relação ao indivíduo ou o “julgamento individual” da partícula. O terceiro termo representa o comportamento social ou global, que indica a parcela de colaboração entre todos os indivíduos (SHI e EBERHART, 1998). As constantes c_1 e c_2 podem ser ajustadas para se colocar mais peso na parte relativa ao indivíduo ou na parte social, respectivamente.

A equação (4.24) ajusta a nova velocidade da partícula em função de três fatores: a) a velocidade da última iteração que pode ser multiplicada por um fator de inércia de forma a regular a extensão do espaço de busca; b) a diferença entre a posição atual e a melhor posição do indivíduo e c) a diferença entre a posição atual da partícula e a melhor posição encontrada entre os elementos do grupo, ou melhor, posição global.

A figura 4.4 indica a condição inicial da última iteração (k), que corresponde a posição inicial (X_i) e a velocidade inicial (V_i). Os valores P_i , P_g e X^* correspondem respectivamente à melhor posição individual, à melhor posição global e à solução do problema ou valor ótimo.

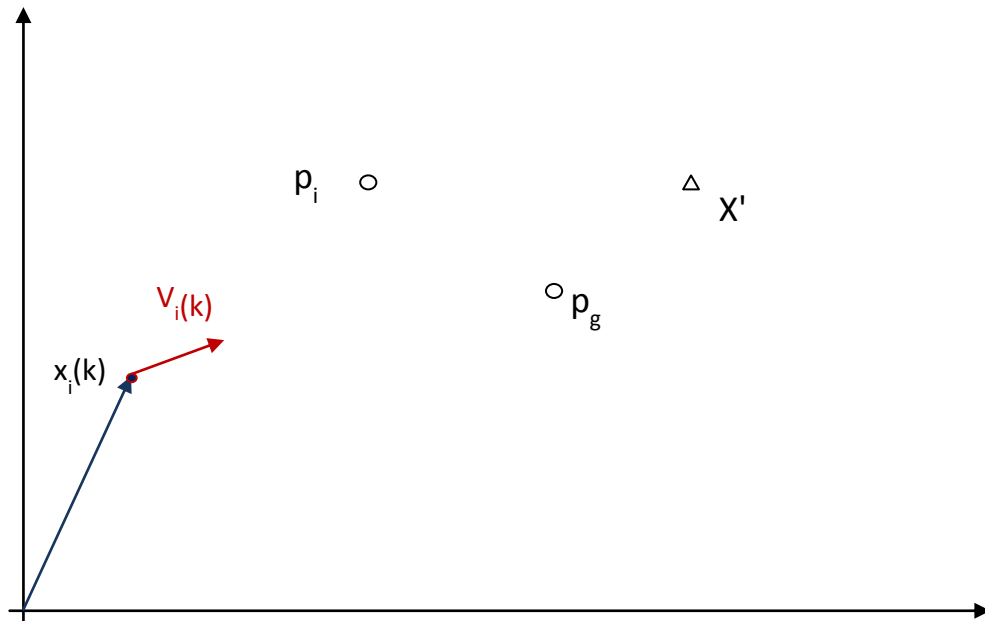


Figura 4.4: Condição inicial para o cálculo da nova posição indicando a última posição e a velocidade da última iteração (k).

A nova posição é ajustada de acordo com a equação (4.25) e depende da posição atual mais a parcela individual da velocidade calculada na equação (4.24). A figura 4.5 indica as três componentes que compõem o cálculo da velocidade indicado na equação (4.24), que correspondem à velocidade da iteração anterior multiplicada pelo fator de inércia e pelos termos relativos à parte cognitiva ou do indivíduo e à parte relativa ao comportamento global. O desempenho da partícula é avaliado em função de uma função de custo previamente determinada e que é função do problema.

A figura 4.6 indica a nova posição da partícula na iteração (k+1) de acordo com a equação (4.25) e a mesma depende da posição atual mais a parcela individual da velocidade calculada na equação (4.24).

Pode-se observar que a nova posição corresponde a uma posição intermediária entre a melhor posição individual e a melhor posição global e essa posição é regulada pelas constantes c_1 e c_2 , cujos valores retratam o compromisso entre o comportamento individual e o social e estão associados a um efeito estocástico em cada componente.

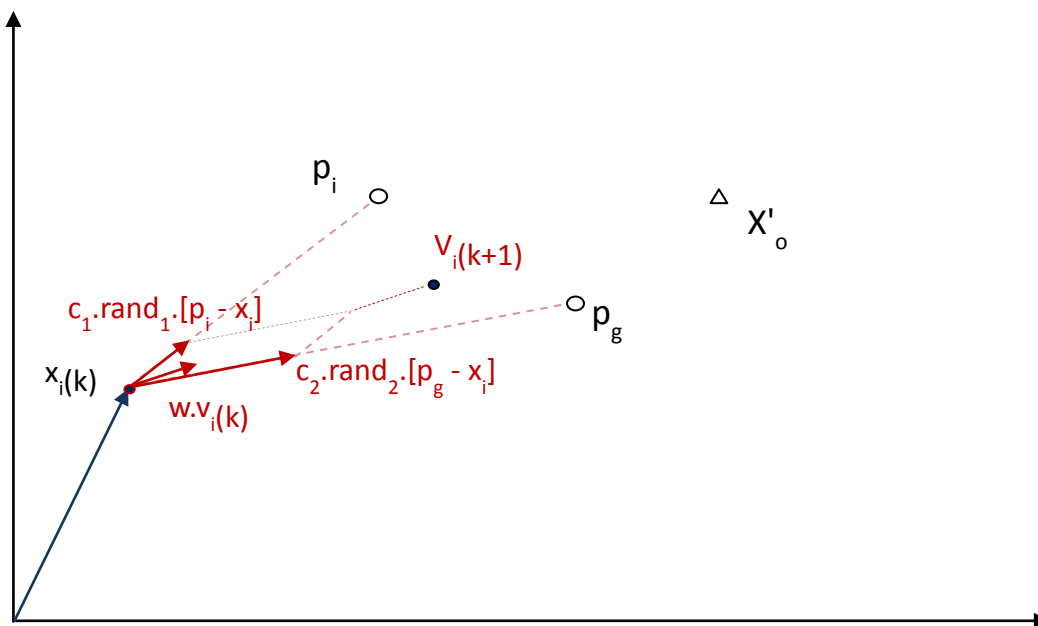


Figura 4.5: Representação Gráfica da Velocidade na iteração (k+1), conforme a equação (4.24)

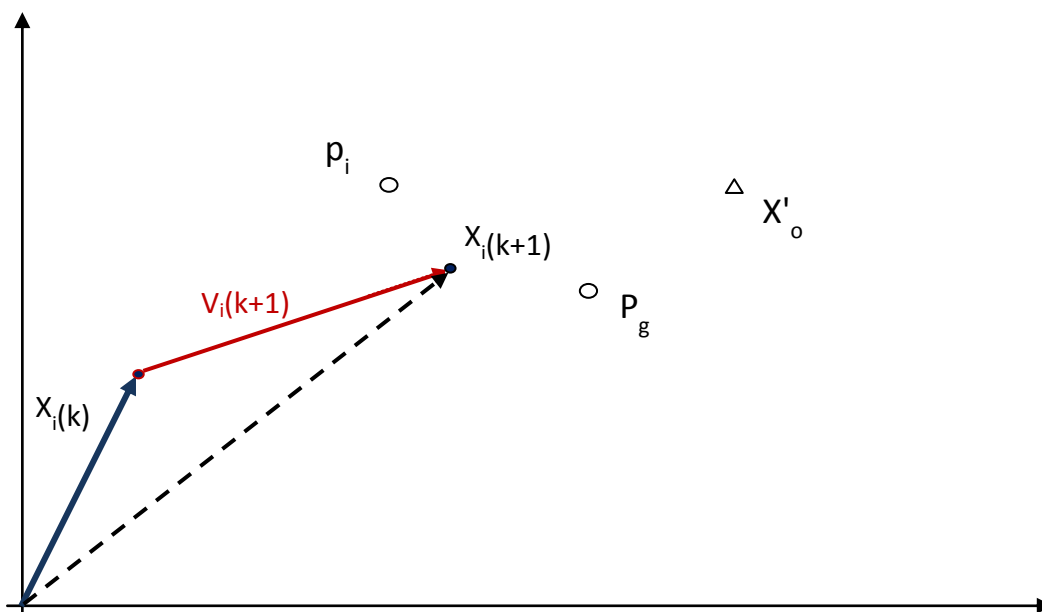


Figura 4.6: Representação gráfica da posição da partícula na iteração (k+1), conforme equação (4.25).

O algoritmo de otimização por enxame de partículas (PSO) mostra-se robusto e eficiente na solução de uma vasta gama de problemas, como problemas não lineares,

não diferenciáveis e multimodais. Assim, o algoritmo PSO parece uma solução natural para a aplicação na Reconciliação de dados ou Reconciliação de dados Robusta.

Nesta tese será utilizado o algoritmo PSO padrão proposto por KENNEDY e EBERHART (1995), da mesma forma que desenvolvido e utilizado em VALDETARO E SCHIRRU (2009, 2011). As constantes de ajuste são aquelas usualmente recomendadas, por exemplo, em BRATON e KENNEDY (2007) e KENNEDY (2007), cujos valores constantes são: $w=0.7298$, $c1=2.05$, $c2=2.05$.

CAPÍTULO 5:

RECONCILIAÇÃO ROBUSTA DE DADOS COM SELEÇÃO DE MODELO DE PROBABILIDADE SIMULTÂNEA

5.1 - Introdução

Conforme foi apresentado no capítulo 4, a etapa de seleção de modelo de probabilidade é uma parte crucial no ajuste de estimadores robustos. Na utilização do estimador de três partes de Hampel ou de outro estimador robusto não há uma função de distribuição de probabilidade definida a priori. Assim, torna-se difícil a realização de testes de hipótese a fim de garantir que uma determinada função de distribuição seja a mais adequada para a amostra em análise.

Em alguns estimadores como o obtido pela função Fair, pode-se relacionar um índice de desempenho, p. ex., a eficiência assintótica, com uma constante de ajuste do estimador, mas esse índice não traz informações precisas sobre a variância do mesmo. Dessa forma, outros índices, como a eficiência relativa, são calculados por meio de simulações e pelo uso do método Monte Carlo e as constantes obtidas, pressupondo que o estimador tem uma eficiência pré-determinada, por exemplo 95%, em relação a uma determinada distribuição (OZYURT e PIKE, 2004). Nos trabalhos de OZYURT e PIKE (2004), BASU e PALIWAL (1989) e PRATA *et al.* (2008) é possível verificar a comparação do desempenho de diversos estimadores, como os estimadores de Andrew, de Bisquare, de Welsch e outros obtidos por meio de simulação e testes baseados nas metodologias indicadas acima.

Outros critérios de seleção de modelos de probabilidade são aqueles baseados em um critério de informação, onde este critério é utilizado como ajuste ou índice para medir qual modelo se ajusta melhor a um conjunto de dados (HAMPEL *et al.*, 1986 e BOZDOGAN, 2000), ou seja, para uma determinada constante do estimador, o valor do

índice de desempenho é calculado, por exemplo o valor do AIC, e aquele modelo de probabilidade do estimador que obtiver o menor índice é o escolhido.

A minimização de um critério de informação não traz informação direta ou específica sobre algum modelo de probabilidade, mas é uma forma adequada de obter o melhor ajuste no senso estatístico para esse modelo (SPANOS, 2010). Na aplicação da estatística robusta, a informação sobre a função de distribuição ou modelo de probabilidade adequado em um primeiro momento não é o principal objetivo, pois parte-se da premissa de que há uma contaminação da amostra e da dificuldade de se obter essa função de distribuição determinada. Assim, a utilização de um critério de informação mostra-se um caminho adequado para a seleção de modelo de probabilidade com estimadores robustos. Deve-se ressaltar que outros resultados sobre a adequação de uma função estatística aplicada ao problema devem ser analisados, para dar maior confiabilidade aos dados estimados.

Diversos critérios são encontrados na literatura e podem-se citar o critério de MALLOW C_p e o Critério de Informação de Akaike, que foi descrito na seção 3.3. Esses métodos são consagrados, eficientes, mas são baseados na hipótese que a distribuição do erro é Normal, com média zero e variância conhecida, o que os torna índices sensíveis ao erro grosseiro, visto que são baseados no estimador de mínimos quadrados (AGOSTINELLI, 2002).

A suposição de que o erro possui distribuição Normal é uma aproximação que nem sempre ocorre na prática. Em relação ao Critério de Informação de Akaike, o mesmo não pode ser utilizado com estimadores robustos (HAMPEL *et al.*, 1986), visto que o mesmo pressupõe a distribuição do erro Normal e a presença de Erros Grosseiros afeta o seu cálculo. Dessa forma, um critério de informação com características mais abrangentes e que possa ser utilizado com estimadores robustos e aplicado à seleção de modelo é desejável.

5.2 – Critério de Informação de Akaike Robusto

RONCHETTI (1985 e 1997a) propôs uma versão Robusta do Critério de Informação de Akaike (AICR), que corresponde exatamente ao estimador de Máxima

Verossimilhança, só que para os M-estimadores, classe essa, da qual os estimadores robustos fazem parte (HAMPEL *et al.*, 1986).

O Critério de Informação de Akaike Robusto (AICR) é definido por

$$AICR(k, \alpha, \rho) = (2) \cdot \sum_{i=1}^m \rho\left(\frac{x_i - \tilde{x}_i}{\hat{\sigma}}, \theta\right) + \alpha \cdot k, \quad (5.1)$$

onde α é constante, ρ é a função que define o M-estimador, θ corresponde as constantes de ajuste do M-estimador, m é o número de observações, x_i são as medidas do processo, \tilde{x}_i correspondem aos valores estimados, $\hat{\sigma}$ advém de uma estimativa robusta da variância (σ^2) e k é o número de parâmetros independentes do estimador (HAMPEL *et al.*, 1986).

No trabalho de RONCHETTI (1985), a relação proposta por AKAIKE (1974) e que corresponde basicamente ao critério de informação de KULLBACK-LIEBLER, que indica a discrepância (“distância”) entre a função densidade de probabilidade verdadeira $f(x|\theta^*)$ e a função densidade que se deseja ajustar $g(x|\theta)$, é definida a partir da equação abaixo, sendo que o AIC é definido na equação (3.28):

$$I(q, p) = 2(q - p)^{-1} \cdot (AIC(p, \alpha) - AIC(q, \alpha)). \quad (5.2)$$

Substituindo-se o Critério de Informação de Akaike, pela versão robusta (AICR) apresentada na equação (5.1) a relação acima adquire a seguinte forma:

$$I(q, p) = 2(q - p)^{-1} \cdot (AICR(p, \alpha) - AICR(q, \alpha)). \quad (5.3)$$

Comparando as equações (5.2) e (5.3) RONCHETTI (1985) mostra que o Critério de Informação de Akaike Robusto (AICR) é o correspondente natural do Critério de Informação de Akaike ao se utilizar estimadores e testes robustos.

Pode-se observar na equação (5.1), que o critério AICR, da mesma forma que o critério AIC, possui dois termos. O primeiro é um termo correspondente ao ajuste dos dados e o segundo corresponde a uma penalidade para evitar o ajuste excessivo ou

solução trivial (“sobreajuste”). O valor do índice AICR corresponderá a uma relação ou compromisso entre o melhor valor de ajuste e a penalização.

Neste trabalho a determinação exata do valor da constante α não será aprofundada, mas a mesma é pouco menor do que 2 de acordo com RONCHETTI (1985, 1997a) e no caso da constante do estimador de Huber, a mesma está entre os valores 1.3 e 1.6.

O objetivo do processo de seleção de modelo de probabilidade baseado no critério de informação de Akaike Robusto (AICR) visa minimizar o índice AICR, conforme indicado na equação (5.1), a fim de determinar o modelo de probabilidade que se ajusta à maioria dos dados medidos, levando em conta que o erro não possui distribuição normal exata.

O valor da constante k para uso na reconciliação de dados é dado pelo número de parâmetros do modelo mais o número de erros grosseiros identificados (n_o), que quando utiliza um estimador redescendente, em especial o estimador de três partes de Hampel, é obtido diretamente pela equação (4.15).

Na equação 5.1, a constante α , o valor estimado da variância $\hat{\sigma}^2$ e o valor da constante k são previamente determinados. Dessa forma, para identificar o melhor M-estimador, é necessário utilizar um algoritmo de otimização global, que nesse trabalho é o algoritmo baseado na inteligência de exames ou PSO, escolhido devido às características positivas já identificadas em trabalhos anteriores de PRATA (2009) e VALDETARO e SCHIRRU (2009). Devido aos resultados promissores, o algoritmo PSO mostrou-se uma alternativa viável e eficiente quando aplicado à reconciliação de dados e identificação de erros grosseiros.

5.3 – Método simultâneo para Reconciliação Robusta de Dados, Identificação de Erros Grosseiros e Seleção de Modelo baseado no Critério de Informação de Akaike Robusto (AICR)

Convém ressaltar que o processo de reconciliação de dados fica mais robusto ao se utilizar um M-estimador no lugar da função objetivo baseada no erro quadrático ou mínimos quadrados ponderados.

Nesse trabalho, o estimador robusto a ser utilizado é o estimador redescendente de três partes de Hampel, apresentado na equação (4.13). Observando-se a formulação geral apresentada na equação (3.1), o problema da reconciliação robusta de dados e identificação de erros grosseiros adquire a seguinte forma:

$$\begin{aligned}
\min_{\tilde{x}, u, p} \sum_{i=1}^n F_H\left(\frac{x_i - \tilde{x}_i}{\hat{\sigma}_i}\right) \quad & s.t. \\
h(\tilde{x}_i, u_i, p) = 0 \quad & x_L \leq \tilde{x}_i \leq x_U \\
g(\tilde{x}_i, u_i, p) \leq 0 \quad & u_L \leq u_i \leq u_U \\
& p_L \leq p \leq p_U,
\end{aligned} \tag{5.4}$$

onde, x_i é o valor medido e \tilde{x}_i o valor estimado da variável, F_H corresponde ao estimador redescendente de três partes de Hampel, $\hat{\sigma}_i$ advém de uma estimativa robusta da variância σ^2 , p é o conjunto de parâmetros, u_i a variável não medida, h o conjunto de restrições de igualdade, g o conjunto de inequações, e os subscritos L e U correspondem aos limites inferiores e superiores de \tilde{x}_i , u_i e p . Considera-se que as constantes do estimador F_H já foram pré-determinadas.

Observando os índices AIC e AICR, ambos são constituídos por duas partes, a primeira parte corresponde ao termo de ajuste e a outra parte a um termo de penalidade, conforme se pode verificar na equação (5.1).

Considerando o problema de reconciliação robusta de dados e identificação de erros grosseiros apresentado na equação (5.4) acima, nota-se que a função objetivo nessa equação corresponde ao primeiro termo ou *termo de ajuste* do Critério de Akaike Robusto apresentado na equação (5.1), e neste trabalho especificamente, o estimador (ρ) é o estimador de três partes de Hampel (F_H).

Nota-se ainda que o segundo termo do AICR (*α.k*) da equação (5.1) está relacionado com a detecção de erros grosseiros, que no caso do estimador redescendente de Hampel está relacionado diretamente com a constante c .

Como mencionado anteriormente, o uso de um estimador robusto sem o devido ajuste das constantes que proporcionam a seleção do modelo de probabilidade pode ocasionar resultados indevidos e a utilização da estratégia de reconciliação de dados robusta, como a apresentada acima, na equação (5.4), pode simplesmente ser uma estratégia ineficiente ou sem efeito.

Entretanto, a minimização do índice AICR em relação às variáveis de ajuste do estimador é uma forma de buscar o melhor ajuste ou um valor ótimo para esses parâmetros, de forma semelhante à minimização do índice AIC apresentado por ARORA e BIEGLER (2001) e WONGRAT *et al.* (2005) e explicado no item 4.5.

Observando as considerações acima, nota-se que a solução do problema de reconciliação robusta de dados apresentado na equação (5.4) corresponde a minimizar o AICR, a menos de um valor constante ($\alpha.k$), dado um valor pré-determinado para as constantes de ajuste do M-estimador, enquanto o processo para determinar os valores das constantes de ajuste do M-estimador corresponde a minimizar o AICR em relação a essas constantes para uma determinada amostra.

Assim, baseado nas características comuns desses dois problemas, propõe-se neste trabalho a utilização de uma nova função objetivo para incorporar à solução do problema da reconciliação de dados e identificação de erros grosseiros a etapa de seleção de modelo de probabilidade ou ajuste das constantes do estimador robusto, de forma que a RD e a IEG e a seleção de modelos sejam feitas de forma simultânea.

Assim, a nova função objetivo proposta aqui para resolver simultaneamente o problema da reconciliação de dados robusta, da identificação de erros grosseiros e da seleção de modelo de probabilidade é formada pelo termo de ajuste do critério de informação AICR, que é o termo comum à formação do índice AICR e a função objetivo do problema de reconciliação robusta de dados apresentado na equação (5.4), somada com o segundo termo do índice AICR, que corresponde à penalidade do referido índice e está associado à detecção de erros grosseiros, além de evitar o ajuste excessivo ou solução trivial (“overfitting”). Dessa forma a nova função objetivo adquire a seguinte forma,

$$\begin{aligned}
& \min_{\tilde{x}, u, p, \theta} 2 \cdot \sum_{i=1}^m F_H \left(\frac{x_i - \tilde{x}_i}{\hat{\sigma}_i}, \theta \right) + \alpha \cdot n_o & \text{s.t.} \\
& h(\tilde{x}_i, u_i, p) = 0 & x_L \leq \tilde{x}_i \leq x_U \\
& g(\tilde{x}_i, u_i, p) \leq 0 & u_L \leq u_i \leq u_U \\
& & p_L \leq p \leq p_U, \\
& & \theta_L \leq \theta \leq \theta_U
\end{aligned} \tag{5.5}$$

onde, x_i é o valor medido e \tilde{x}_i é o valor estimado da variável, θ corresponde as constantes de ajuste do M-estimador, aqui, F_H corresponde ao estimador redescendente de três partes de Hampel, $\hat{\sigma}_i$ advém de uma estimativa robusta da variância σ^2 de cada componente, p é o conjunto de parâmetros, u_i a variável não medida, h o conjunto de restrições de igualdade, g o conjunto de inequações, e os subscritos L e U correspondem aos limites inferiores e superiores de \tilde{x}_i, u_i, p e θ .

Nota-se que a estratégia apresentada no método proposto acima corresponde a minimizar o Critério de Informação de Akaike Robusto diretamente, otimizando o valor das constantes de ajuste do estimador e ao mesmo tempo resolvendo o problema da Reconciliação de Dados e Identificação de Erros Grosseiros.

Deve-se observar que no problema de reconciliação de dados apresentado na equação (5.4), o valor do parâmetro θ é fixo, mas na nova função objetivo proposta aqui nesse trabalho, os parâmetros de ajuste (θ) correspondem as constantes de ajuste do M-estimador, que agora, nesse problema, são tratadas como variáveis.

Na formulação apresentada na equação (5.5), o termo relacionado à penalidade depende da determinação do número de erros grosseiros (n_o), que no caso do estimador redescendente de três partes de Hampel pode ser calculado por meio da expressão apresentada na equação (4.15), que indica o ponto de corte. Caso o valor do resíduo seja maior do que a constante de ajuste c há a indicação de um erro grosseiro.

Outro aspecto é como deve ser considerada a estimativa das três constantes (a, b e c) do estimador redescendente de três partes de Hampel. No caso de uma busca multidimensional, o estimador pode perder a característica de resistência e robustez. ARORA e BIEGLER (2001) sugerem manter uma relação de proporcionalidade entre

essas constantes, como indicado na equação (5.6). Essa relação de proporcionalidade não é considerada ótima, mas pode fornecer um bom ajuste devido a sua pouca influência na função objetivo (ARORA e BIEGLER, 2001).

$$c \geq b + 2a \quad \text{onde} \quad b = \frac{c}{2} \quad e \quad a = \frac{(c-b)}{2} = \frac{c}{4} \quad , \quad (5.6)$$

Uma vez estabelecidas as condições relativas às constantes de ajuste, o problema de reconciliação robusta de dados utilizando o estimador de três partes de Hampel, identificação de erros grosseiros e seleção de modelo simultânea assume a seguinte forma,

$$\begin{aligned} \min_{\tilde{x}, u, c} \quad & 2 \cdot \sum_{i=1}^n F_H\left(\frac{x_i - \tilde{x}_i}{\hat{\sigma}_i}, c\right) + 1.9 \cdot n_o \quad s.t. \\ h(\tilde{x}_i, u_i) = 0 \quad & x_L \leq \tilde{x}_i \leq x_U \\ g(\tilde{x}_i, u_i) \leq 0 \quad & u_L \leq u_i \leq u_U \\ b = c/2 \quad & c_L \leq c \\ a = c/4 \end{aligned} \quad (5.7)$$

onde, x_i é o valor medido e \tilde{x}_i é o valor estimado da variável, c é o valor estimado da constante de ajuste do estimador robusto, F_H corresponde ao estimador redescendente de três partes de Hampel, $\hat{\sigma}_i$ advém de uma estimativa robusta da variância σ^2 , u_i é a variável não medida, h é o conjunto de restrições de igualdade, g é o conjunto de inequações, e os subscritos L e U correspondem aos limites inferiores e superiores de \tilde{x}_i , u_i e c .

Deve-se ressaltar que a constante c corresponde à constante de ajuste do M-estimador e aqui a mesma é tratada como variável de decisão. Para o problema de RD e IEG, n_o é dado pelo número de parâmetros mais o número de erros grosseiros identificados. A constante a foi escolhida com um valor pouco menor do que 2,

conforme indicado no trabalho de HAMPEL et. al. (1986) e o valor utilizado foi igual a 1,9.

Ao resolver o problema acima utilizando o algoritmo PSO, a solução ótima é o vetor solução correspondente à melhor posição global entre as diversas partículas. O vetor solução consiste em n variáveis estimadas mais a variável de ajuste do estimador c , resultando em um vetor solução com n+1 variáveis, como indicado abaixo:

$$\mathbf{x}_{n+1} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots, \tilde{x}_n, c) \quad (5.8)$$

A utilização da nova função objetivo proposta neste trabalho e indicada na equação (5.5) e detalhada na equação (5.7) permite resolver três problemas simultaneamente: o primeiro é obter uma solução para o problema da reconciliação robusta de dados; o segundo é a seleção do modelo probabilístico que melhor se ajusta aos dados estatísticos, assim obtendo o melhor ajuste das constantes do estimador e o terceiro é a identificação da ocorrência de erros grosseiros diretamente. Dessa forma, para efeito de identificação desse novo método, neste trabalho o denominaremos como método de **R**econciliação Robusta de **D**ados com **S**eleção de **M**odelo **S**imultânea (RDSMS).

5.4 – Considerações sobre o Método de Reconciliação Robusta de Dados com Seleção de Modelo Simultânea (RDSMS)

O método RDSMS acrescentou uma melhoria em relação a métodos anteriores com a introdução da etapa de ajuste das constantes do estimador de forma simultânea, eliminando uma etapa prévia de ajuste como apresentada no método proposto por ARORA e BIEGLER (2001) e WONGRAT *et al.* (2005).

Considerando que outros métodos de reconciliação de dados acabam por utilizar simulações para calcular índices de desempenho, como o cálculo da eficiência relativa,

a qual é determinada por meio de simulações e uso do método Monte Carlo, o método RDSMS aqui apresentado também mostra vantagens de uso sobre esses métodos (seção 5.1), visto que o cálculo é feito de forma exata determinando o conjunto principal de dados e minimizando os índices de desempenho do estimador robusto como a sensibilidade a erros grosseiros e diminuindo o desvio assintótico (“bias”).

Outra vantagem é a facilidade de implementação do método RDSMS e também o uso de um algoritmo de otimização global eficiente, que é o PSO, implementado de forma padrão. Como a função objetivo depende de um horizonte de medidas para realizar a estimação robusta dos dados, o método RDSMS pode ser utilizado de forma dinâmica utilizando uma janela de dados de tamanho h , que se desloca a cada iteração. Dessa forma, se a característica da estatística dos dados sofrer alguma alteração, o método RDSMS se auto-ajusta, calculando as constantes do estimador a cada iteração.

Convém observar que o método RDSMS pode ser aplicado utilizando outros estimadores, mas esse assunto deve ser ainda pesquisado em mais detalhes e nesta tese não serão feitas quaisquer avaliações ou comparações específicas entre diversos estimadores a fim de determinar o melhor tipo de estimador robusto a ser utilizado. Aqui consideramos que o estimador robusto de três partes de Hampel possui excelentes características, as quais já foram apresentadas anteriormente, e que superam parte considerável dos estimadores robustos.

Outra questão em aberto e não abordada neste trabalho, é a relação entre as constantes de ajuste do estimador de Hampel e a proporcionalidade entre elas, que pode ser explorada a fim de se obter proporções entre elas que atendam melhores índices de robustez, como a sensibilidade a desvios na medida (λ).

No próximo capítulo, serão apresentados os resultados de exemplos simulados e aplicações em cenário real a fim de mostrar as características efetivas do método de Reconciliação de Dados Robusta com Seleção de Modelo Simultânea (RDSMS). Serão apresentados os resultados de dois exemplos simulados e um exemplo com dados reais da usina de Angra 2 que utilizam a técnica apresentada nesta tese, com o intuito de avaliar o desempenho do método proposto.

CAPÍTULO 6:

RESULTADOS

6.1 - Introdução

Neste capítulo serão apresentados os resultados obtidos com a implementação do método RDSMS.

Para fins de avaliação do método proposto, inicialmente será feito um ajuste não simultâneo, baseado na estratégia proposta por ARORA e BIEGLER (2001), mas utilizando o critério de informação robusto (AICR), determinando assim, o valor ótimo das constantes a , b , e c do estimador de três partes de Hampel para o problema considerado.

Uma vez determinados os valores ótimos ou quase ótimos das constantes de ajuste, será feita a comparação desses valores ótimos com aqueles obtidos com o método RDSMS proposto neste trabalho.

Dessa mesma forma serão analisados dois exemplos. O primeiro corresponde a um sistema não linear, que é bastante utilizado como “benchmark” por diversos autores para avaliar o desempenho de vários métodos (PAI e FISHER, 1988). O segundo corresponde a uma realização do balanço de massa simplificado de uma turbina a vapor de um circuito secundário de uma usina nuclear (NPP) típica, o qual foi baseado na norma VDI-2048 (2000) e usado aqui com dados simulados no cálculo da potência térmica do reator.

O terceiro exemplo corresponde a uma realização de um balanço de massa simplificado de uma turbina a vapor do circuito secundário da usina nuclear de Angra 2, que utiliza dados reais obtidos do processo. Estes dados serão utilizados “off-line”, ou seja, os dados da planta são gravados em intervalos regulares em um arquivo e após a gravação dos mesmos, o arquivo gerado é utilizado como entrada para o método desenvolvido nesse trabalho. O objetivo principal desta aplicação é avaliar apenas o método proposto neste trabalho em um cenário realista, embora simplificado.

Convém ressaltar que os dados obtidos da usina de Angra 2 são uma amostragem do processo e não são aqueles apropriados para qualquer avaliação ou inferência sobre as condições da usina ou seu desempenho.

6.2 – Exemplo Não linear (PAI e FISHER, 1988)

Este exemplo considera a aplicação da reconciliação de dados e identificação de erros grosseiros a um sistema não linear extraído do trabalho de PAI e FISHER (1988) e testado também no trabalho de WONGRAT *et al.* (2005) com um algoritmo genético modificado, em ARORA e BIEGLER (2001) com um estimador redescendente e usando um método de otimização convencional, em TJOA e BIEGLER (1991), ZHOU *et al.* (2006), VALDETARO e SCHIRRU (2009) e em diversos outros trabalhos. Neste exemplo será apresentado o resultado da aplicação do método automático proposto nesta tese (item 6.3) e também apresentado em VALDETARO e SCHIRRU (2011a e b).

O Problema possui cinco variáveis medidas (x_i) e três variáveis não medidas (u_i) e seis restrições não lineares, a saber:

$$2x_1 + x_2 \cdot x_3 \cdot u_1 + u_2 - u_3 - 126.6 = 0 \quad (6.1a)$$

$$0.5x_1^2 - 0.7x_2 + x_3 \cdot u_1 + x_2^2 \cdot u_1 \cdot u_2 + 2x_3 \cdot u_3^2 - 255.8 = 0 \quad (6.1b)$$

$$x_1 - 2x_2 + 3x_1 \cdot x_3 - 2x_2 u_1 - x_2 \cdot u_2 \cdot u_3 + 111.2 = 0 \quad (6.1c)$$

$$x_3 \cdot u_1 - x_1 + 3x_2 + x_1 \cdot u_2 - x_3 \cdot u_3^{0.5} - 33.57 = 0 \quad (6.1d)$$

$$x_4 - x_1 - x_3^2 + u_2 + 3u_3 = 0 \quad (6.1e)$$

$$x_5 - 2x_3 \cdot u_2 \cdot u_3 = 0 \quad (6.1f)$$

O sistema acima possui a seguinte solução exata:

$$x_e = [4.5124, 5.5819, 1.9260, 1.4560, 4.8545] \quad (6.2a)$$

$$u_e = [11.070, 0.61467, 2.0504] \quad (6.2b)$$

As cinco variáveis medidas são simuladas de acordo com a equação (6.3), onde o valor de x_i é simulado adicionando-se o ruído (η) com desvio padrão $\sigma = 0.1$ e erro grosseiro (t) correspondente a 25σ .

$$x_i = x_e + \eta + t \quad (6.3)$$

Para as variáveis não medidas (u_i), utiliza-se a técnica proposta por PRAKOTPOL e SRINOPHAKUN (2003) para a solução do sistema de equações e obtenção de uma solução viável. Dividem-se as variáveis do problema em dois grupos: variáveis medidas, x_i , e variáveis não medidas. Os valores das medidas são aqueles gerados pelo algoritmo de otimização, que aqui é o algoritmo baseado na inteligência de enxames de partículas. As variáveis não medidas são calculadas em função dos valores das variáveis medidas por meio da solução simultânea do sistema de equação, que fica reduzido, após substituir o valor de cada variável independente pelo valor gerado pelo PSO.

Neste exemplo foram gerados 100 valores para cada componente x_i de acordo com a equação (6.3), que corresponde ao horizonte de medidas h igual a 100. Esses dados correspondem as variáveis medidas e simulam dados adquiridos em tempo real.

Cada componente das variáveis medidas foram corrompidas por 20 erros grosseiros cada, ou seja, os 20 primeiros erros grosseiros foram adicionados às 20 primeiras medidas da variável x_1 , os próximos 20 erros grosseiros foram adicionados à variável x_2 após as primeiras 20 medidas de x_2 e assim por diante até a componente x_5 e completar 100 erros grosseiros.

O algoritmo foi implementado para trabalhar como uma janela dinâmica com horizonte de tamanho H cujo valor de H é igual a 100, sendo que, a cada ciclo, um novo valor é lido e o valor mais antigo é descartado, mantendo-se o horizonte de medidas constante.

Como a janela dinâmica se desloca uma medida a cada ciclo, para que ela possa se deslocar, foram acrescentadas 100 medidas iguais às geradas inicialmente para cada componente, totalizando 200 medidas, assim, em um determinado instante t, o estimador utiliza no cálculo da f_0 as 100 medidas do horizonte de tempo em cada componente desde o instante t-100 e identifica o erro grosseiro no instante t.

O mesmo padrão de ocorrência dos erros grosseiros foi mantido para efeito de simplificação e o objetivo é avaliar a robustez do estimador na presença dos diversos erros grosseiros gerados e também a capacidade do estimador identificar corretamente em quais variáveis medidas ocorreram erros grosseiros no ciclo corrente.

A função objetivo para o problema proposto reflete o apresentado na equação (5.7) e está indicado abaixo.

$$\min_{x,\mu,c} 2 \cdot \sum_{i=1}^5 \sum_{j=1}^{H=100} F_H(x_{ij} - \tilde{x}_i, c) + 1.9n_o, \quad (6.4)$$

onde F_H corresponde ao modelo de distribuição do estimador redescendente de três partes de Hampel conforme a equação (4.13) e as restrições do problema são dadas pelas equações (6.1a-f) e a relação entre as constantes obedecem a aquelas indicadas na equação (5.6).

Este exemplo foi inicialmente utilizado para avaliar o comportamento e desempenho do algoritmo baseado em enxame de partículas ao se utilizar na função objetivo o estimador redescendente de três partes de Hampel (4.13). Além disso, o trabalho serviu para verificar a forma de ajuste das constantes conforme proposto no trabalho de WONGRAT *et al.* (2005).

Uma avaliação mais detalhada sobre o comportamento e desempenho do algoritmo PSO aplicado à reconciliação robusta de dados e identificação de erros grosseiros com o uso do estimador redescendente de Hampel pode ser consultada no trabalho de VALDETARO e SCHIRRU (2009) e na mesma época nos trabalhos de PRATA (2009) e PRATA *et al.* (2009) mencionando o uso do algoritmo PSO na reconciliação robusta de dados com o estimador de Welsch e estimador quadrático, respectivamente.

O algoritmo PSO implementado foi o PSO padrão com os seguintes parâmetros usuais e constantes: $w=0.7298$, $c_1=2.05$, $c_2=2.05$. Foram utilizadas 60 partículas (n_p) e cerca de 120 iterações (n_i) a cada passo.

Baseado nos efetivos resultados obtidos no ajuste “off-line” em ARORA e BIEGLER (2001), WONGRAT *et al.* (2005) e VALDETARO e SCHIRRU (2009), a metodologia para validar o procedimento de reconciliação de dados robusta, identificação de erros grosseiros e seleção de modelo simultâneas foi a de comparar o resultado obtido com este procedimento, com o resultado da estratégia de ajuste manual ou “off-line” como apresentada no item 4.5, utilizando o índice de desempenho AICR ao invés do índice AIC.

A tabela 6.1 apresenta os resultados do ajuste feito manualmente ou “off-line”, que estão apresentados nas linhas R1 a R8. Em cada linha, nas colunas de x_1 a x_5 e de u_1 a u_3 , estão indicadas a melhor posição global entre as partículas (p_g), n_o é o número de erros grosseiros, AICR é o Critério de Informação de Akaike Robusto, X_e corresponde a solução exata.

Na tabela 6.1 abaixo, em relação ao ajuste realizado separadamente (“off-line”), pode-se determinar que o valor mínimo do AICR encontra-se entre os valores R4 e R6 e há indicação de que o mesmo esteja muito próximo ao valor R5. Pode-se notar que para os ajustes a partir de R5 até R8 o valor de AICR aumenta devido à perda na qualidade da estimação e aumento do número de erros grosseiros detectados. Em relação aos valores R1 e R2 o aumento do índice AICR aumenta devido à influência do termo de ajuste que é mais significativo do que o número de erros grosseiros e indica também uma perda na qualidade da estimação.

Tabela 6.1 - Resultados do ajuste “off-line” do estimador redescendente (R1-R8) e valores calculados, obtidos com o método RDSMS (X_{opt}).

| Estim. | <i>M-estimator constants</i> | AICR | n_o | x_1 | x_2 | x_3 | x_4 | x_5 | u_1 | u_2 | u_3 |
|-------------|--|---------------|------------|----------------|----------------|----------------|----------------|----------------|------------------|----------------|----------------|
| R1 | $a=1.0000$ $b=2.0000$ $c=4.0000$ | 1.1323 | 60 | 4.52221 | 5.57111 | 2.07957 | 1.72065 | 4.84090 | 10.29088 | 0.52931 | 2.19895 |
| R2 | $a=0.5000$ $b=1.0000$ $c=2.0000$ | 0.5682 | 80 | 4.50605 | 5.58037 | 1.92930 | 1.64000 | 4.86291 | 11.04728 | 0.63510 | 1.98438 |
| R3 | $a=0.2500$ $b=0.5000$ $c=1.0000$ | 0.4500 | 100 | 4.52054 | 5.59407 | 1.91936 | 1.45141 | 4.86239 | 11.08165 | 0.61955 | 2.04451 |
| R4 | $a=0.1250$ $b=0.2500$ $c=0.5000$ | 0.4026 | 100 | 4.50284 | 5.58419 | 1.90489 | 1.45814 | 4.84107 | 11.18495 | 0.63088 | 2.01415 |
| R5 | $a=0.0625$ $b=0.1250$ $c=0.2500$ | 0.4002 | 103 | 4.49532 | 5.58051 | 1.93714 | 1.46828 | 4.84715 | 11.01347 | 0.60818 | 2.05713 |
| R6 | $a=0.0500$ $b=0.1000$ $c=0.2000$ | 0.4282 | 111 | 4.55345 | 5.60146 | 1.94737 | 1.47250 | 4.86012 | 10.90832 | 0.59643 | 2.09225 |
| R7 | $a=0.0375$ $b=0.0750$ $c=0.1500$ | 0.5554 | 145 | 4.49893 | 5.58556 | 1.88813 | 1.46780 | 4.86670 | 11.27735 | 0.65024 | 1.98197 |
| R8 | $a=0.0313$ $b=0.0625$ $c=0.1250$ | 0.6492 | 170 | 4.51600 | 5.57191 | 1.92383 | 1.43633 | 4.81638 | 11.10296 | 0.60841 | 2.05746 |
| Xe | | | | 4.51240 | 5.58190 | 1.92600 | 1.45600 | 4.85450 | 11.07000 | 0.61467 | 2.05040 |
| Xopt | $a=0.0612$ $b=0.1223$ $c=0.2446$ | 0.3884 | 100 | 4.50369 | 5.57453 | 1.92224 | 1.43968 | 4.84268 | 11.107695 | 0.61507 | 2.04799 |
| Eopt | | | | 0,00871 | 0,00737 | 0,00376 | 0,01632 | 0,01182 | 0,037695 | 0,00040 | 0,00241 |

Após determinar os valores ótimos ou quase ótimos para o ajuste manual ou “off-line”, que se situa entre os ajustes R4 e R6, o próximo passo foi obter o valor ótimo das variáveis x_1 a x_5 e u_1 a u_3 aplicando o método RDSMS, ou método automático, e avaliar seu comportamento e desempenho.

O resultado obtido com a aplicação do método RDSMS está mostrado também na tabela 6.1, na linha indicada por X_{opt} e corresponde ao valor esperado e obtido utilizando o método “off-line”, ou seja, entre os valores R4 e R6 e muito perto do valor R5 ($c=0.2446$). Observa-se ainda que na solução X_{opt} , o número de erros grosseiros foi estimado corretamente.

Existe ainda uma pequena diferença entre os resultados do ajuste “off-line” e do método automático, que é atribuída a pouca precisão do método de busca utilizado no ajuste não simultâneo.

A figura 6.1 apresenta o resultado do ajuste automático, mostrando um gráfico indicando o comportamento do índice AICR versus a constante de ajuste do estimador robusto (c) baseado nos valores R1 a R8 da tabela 6.1 e o valor de X_{opt} , que é o valor da constante c obtido pelo método RDSMS e está indicado na tabela 7.1 (0.2446, 0.3884).

O resultado obtido com o método RDSMS indica sua eficiência ao detectar o número correto de erros grosseiros em cada variável e que os valores estimados estão

muito próximos dos valores exatos, indicados na linha X_e (ver tabela 6.1, $E_{opt} = \text{Abs}(X_e - X_{opt})$).

Todos os erros grosseiros foram identificados, indicando uma eficiência de 100% na detecção considerando o teste proposto, onde o erro grosseiro simulava o comportamento de um sensor travado em fundo de escala.

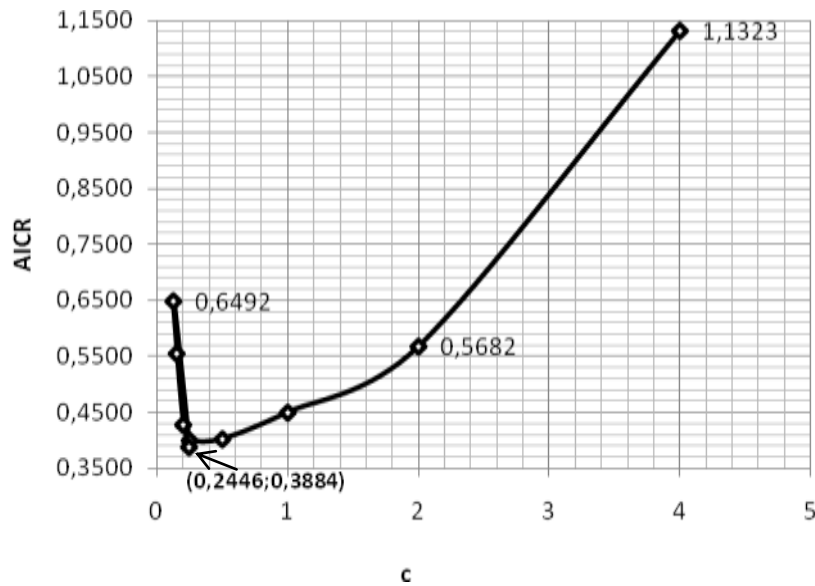


Figura 6.1 – Gráfico AICR x constante de ajuste c

Os testes foram realizados em um computador Core 2 Duo, 1.6 GHz, e cada passo teve um valor de duração médio de 175 segundos.

6.3 – Exemplo de Cálculo da Potência do Reator baseado na norma VDI-2048.

Este exemplo considera a aplicação do método simultâneo para a reconciliação de dados, identificação de erros grosseiros e seleção de modelo de probabilidade ao cálculo da potência térmica de um reator nuclear do tipo PWR, baseado na norma VDI-2048 (2000), cujo diagrama foi apresentado no capítulo 2 e apresentado em VALDETARO e SCHIRRU (2011a).

A utilização do exemplo baseado na norma VDI-2048 tem como objetivo permitir a comparação e a avaliação de desempenho entre o método clássico e o método automático proposto neste trabalho.

O Diagrama Simplificado do Circuito Secundário de uma Usina Nuclear tipo PWR foi apresentado na figura 2.2 e de acordo com o exemplo da norma VDI-2048 (2000), o balanço de massa está descrito na seção 2.3.

As restrições do processo ou os balanços de massa foram determinados nas equações (2.6) a (2.8) e um vetor \mathbf{x} de variáveis medidas foi formado para a aplicação do método de reconciliação de dados:

$$\mathbf{X} = \begin{cases} [m_{GV1}, m_{GV2}, m_{ag1}, m_{fag2}, m_V, m_C, m_{A7}, m_{A6}, m_{A5}, m_{HPC}, m_T] \\ ou \\ [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}] \end{cases} \quad (6.5)$$

Nesse exemplo, os valores exatos das variáveis indicadas na norma VDI-2048 (2000) foram corrompidos por erros aleatórios, sendo que, as variáveis de x_1 a x_4 também foram corrompidas por 20 erros grosseiros cada. Nestas componentes foram adicionados 20 erros grosseiros a x_1 , mais 20 erros grosseiros foram adicionados a x_2 e assim por diante, até a componente x_4 , totalizando-se 80 erros grosseiros.

Da mesma forma que no item anterior, o algoritmo foi implementado para trabalhar como uma janela dinâmica com horizonte de tamanho H cujo valor neste exemplo é igual a 100, sendo que, a cada novo valor lido, o valor mais antigo é descartado, mantendo-se o horizonte de medidas constante.

Como a janela dinâmica se desloca uma medida a cada ciclo, para que ela possa se deslocar, foram acrescentadas 100 medidas iguais às geradas inicialmente para cada componente, totalizando 200 medidas, assim, em um determinado instante t , o estimador utiliza no cálculo da f_o as 100 medidas do horizonte de tempo anteriores ao instante t e identifica o erro grosseiro no instante atual.

O mesmo padrão de ocorrência dos erros grosseiros foi mantido para efeito de simplificação e o objetivo é avaliar a robustez do estimador na presença dos diversos

erros grosseiros gerados nos instantes passados e também a capacidade do estimador identificar corretamente em quais variáveis medidas ocorreram erros grosseiros no ciclo corrente.

As quatro variáveis medidas foram simuladas de acordo com a equação (6.3), onde o valor de x_i é simulado adicionando-se o valor exato mais o ruído (η) com desvio padrão $\sigma = 0.1$ e mais o erro grosseiro (ι) correspondente a 25σ .

O algoritmo PSO implementado foi o PSO padrão com os seguintes parâmetros usuais e constantes: $w=0.7298$, $c_1=2.05$, $c_2=2.05$. Foram utilizadas 120 partículas (n_p) e cerca de 160 iterações (n_i) a cada passo.

O primeiro passo foi realizar um ajuste “off-line” ou manualmente, de forma semelhante ao proposto em ARORA e BIEGLER (2001) e WONGRAT *et al.* (2005), mas ao invés de utilizar o Critério de Informação de Akaike, foi utilizado o AICR, pois o estimador utilizado é um estimador robusto.

A tabela 6.2 apresenta os resultados obtidos com o ajuste “off-line” (R1-R9) e entre esses valores de AICR obtidos, pode-se localizar um valor mínimo entre os resultados R6 e R7, para $c=0.5$ e $c=0.25$, respectivamente.

Tabela 6.2 - Resultados do ajuste “off-line” do estimador redescendente (R1-R9) e valores calculados obtidos com o método RDSMS (X_{opt}).

| Rx(a.b.c.AICR.no) | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | x11 |
|--|----------------|----------------|----------------|----------------|---------------|----------------|----------------|---------------|---------------|----------------|---------------|
| R1 (2; 4; 8; 0.37512; 0) | 45.1862 | 44.6115 | 45.1368 | 44.8816 | 0.5348 | 70.0126 | 10.3533 | 3.7384 | 4.3986 | 18.5083 | 2.0976 |
| R2 (1; 2; 4; 0.28038; 0) | 44.9210 | 44.3365 | 44.8528 | 44.6015 | 0.5191 | 69.9958 | 10.3587 | 3.7215 | 4.4082 | 18.4986 | 2.0999 |
| R3 (0.75; 1.5; 3; 0.20336; 0) | 44.7525 | 44.2172 | 44.7251 | 44.4697 | 0.4985 | 69.9877 | 10.3540 | 3.7468 | 4.3811 | 18.4895 | 2.0957 |
| R4 (0.5; 1; 2; 0.16894; 0) | 45.1960 | 44.6231 | 45.1430 | 44.8861 | 0.5157 | 69.9999 | 10.3757 | 3.7503 | 4.3937 | 18.5068 | 2.0854 |
| R5 (0.25; 0.5; 1; 0.17050; 80) | 44.7032 | 44.1286 | 44.6422 | 44.3863 | 0.5184 | 70.0076 | 10.3643 | 3.7394 | 4.3796 | 18.5002 | 2.1046 |
| R6 (0.125; 0.25; 0.5; 0.15231; 80) | 44.7197 | 44.1352 | 44.6426 | 44.3821 | 0.5296 | 70.0022 | 10.3576 | 3.7533 | 4.3790 | 18.4994 | 2.0935 |
| R7 (0.0625; 0.125; 0.25; 0.16099; 80) | 44.7029 | 44.1147 | 44.6301 | 44.3677 | 0.5343 | 70.0087 | 10.3733 | 3.7625 | 4.4269 | 18.4896 | 2.1249 |
| R8(0.03125; 0.0625; 0.125; 0.63194; 364) | 44.7225 | 44.1346 | 44.6296 | 44.3780 | 0.5130 | 111.2246 | 10.3557 | 3.7278 | 4.4438 | 18.4933 | 2.0760 |
| R9(0.025; 0.05; 0.100; 0.75710; 437) | 44.6629 | 13.2397 | 44.6650 | 44.3832 | 0.5031 | 70.0097 | 10.3483 | 3.7549 | 4.3665 | 18.5252 | 2.1005 |
| Xe | 44.6960 | 44.1230 | 44.6430 | 44.3860 | 0.5240 | 70.0050 | 10.3640 | 3.7440 | 4.3910 | 18.4990 | 2.0920 |
| Xopt(0.0766; 0.1532; 0.3063; 0.14701; 80) | 44.6644 | 44.1245 | 44.6362 | 44.3806 | 0.5323 | 70.0095 | 10.3711 | 3.7503 | 4.3721 | 18.4995 | 2.0793 |
| Eopt | 0.0316 | 0.0015 | 0.0068 | 0.0054 | 0.0083 | 0.0045 | 0.0071 | 0.0063 | 0.0189 | 0.0005 | 0.0127 |

Da mesma forma que no exemplo anterior apresentado no item 6.2, o próximo passo foi obter o valor ótimo das variáveis x_1 a x_{11} aplicando o método RDSMS e avaliar seu comportamento e desempenho.

O resultado do ajuste com o método RDSMS está apresentado na linha X_{opt} da tabela 6.2 e corresponde a um valor ótimo com a constante de ajuste $c=0.3063$. Este valor está dentro do intervalo previsto no ajuste “off-line” realizado na primeira parte deste exemplo, ou seja, entre as medidas R6 e R7 da tabela 6.2.

O resultado obtido com o método RDSMS destaca sua eficiência ao detectar o número correto de erros grosseiros em cada variável, que são 80 erros grosseiros no total, com os valores estimados (X_{opt}) muito próximos dos valores exatos, indicados na linha X_e (ver tabela 6.2, $E_{opt} = \text{Abs}(X_e - X_{opt})$).

Todos os erros grosseiros foram identificados, indicando uma eficiência de 100% na detecção considerando o teste proposto. O erro grosseiro simulava o comportamento de um sensor travado em fundo de escala.

A figura 6.2 apresenta um gráfico indicando o comportamento do índice AICR versus a constante de ajuste do estimador robusto (c) baseado nos valores R1 a R9 da tabela 6.2 e o valor de X_{opt} calculado pelo método RDSMS indicado na tabela 6.2 (0.3063, 0.14701).

É importante ressaltar que o método RDSMS proposto neste trabalho mostrou comportamento semelhante nos dois exemplos apresentados nos itens 6.2 e 6.3, sendo que os resultados calculados por ele (x_{opt}) se situaram entre os valores previstos no método de ajuste em separado (“off-line”) e a quantidade de erros grosseiros identificada foi a correta.

Na próxima seção serão feitas considerações relativas à aplicação do método RDSMS ao cálculo da potência térmica do reator do tipo PWR, baseado no exemplo extraído da norma VDI 2048 (2000) e exemplificada nesta seção e no item 2.3.

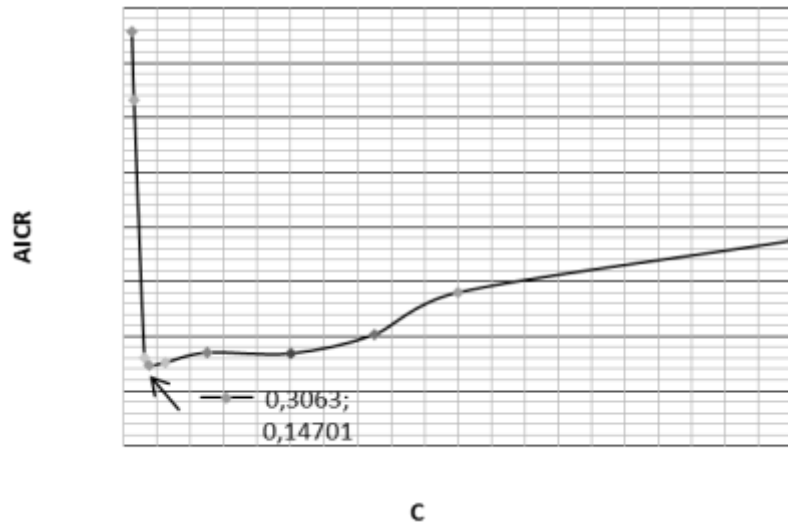


Figura 6.2 – Gráfico AICR x constante de ajuste c

6.3.1 – Considerações sobre o cálculo da potência térmica do reator

A potência térmica do reator (P_R) pode ser calculada a partir da determinação da carga térmica do Gerador de Vapor (\dot{Q}_{GV}), que é proporcional à vazão total de água de alimentação ($m_{ag} = m_{ag1} + m_{ag2}$) é calculada pela equação (2.2) e reapresentada abaixo para efeito de maior clareza na exposição,

$$\dot{Q}_{GV} = (h_s - h_e) \cdot m_{ag} \quad , \quad (6.6)$$

onde h_s é a entalpia do fluido na saída do Gerador de Vapor (GV), h_e é a entalpia do fluido na entrada do GV..

O valor da entalpia do fluido depende da temperatura e pressão do meio, ou seja, na entrada e na saída do GV, que em condições estáveis ou em regime permanente pode-se assumir que são constantes. Como os processos de medição de pressão e temperatura possuem boa precisão e estas variáveis são constantes, pode-se assumir que a medida da entalpia também é constante e conhecida com boa precisão.

Assim, a propagação de erro no cálculo da carga térmica do GV depende da medição da vazão de água de alimentação. Entretanto, o processo de medição de vazão possui uma incerteza significativa, sendo que a precisão dependendo do método de medida pode variar de 0,5% a 2% e em alguns casos pode chegar a 5% de erro (ANDRADE *et al.*, 2002).

Baseado no procedimento apresentado no capítulo 2 e rerepresentado simplificadamente acima, verifica-se que a propagação do erro no cálculo da carga térmica do GV depende fortemente e diretamente do erro de medição do fluxo de massa de água de alimentação.

Utilizando-se o valor reconciliado ao invés do valor medido do fluxo de água de alimentação, verifica-se a diminuição da incerteza na medição, a qual é apresentada na tabela 6.3.

Tabela 6.3 – Cálculo da carga térmica do GV no exemplo simplificado do Circuito Secundário de uma Usina Nuclear tipo PWR.

| R | x3 | x4 | m_{ag} | Qsg norm | Qsg error | Error (MWt) |
|-------------|-----------------|-----------------|----------|----------|-----------|-------------|
| R1 | 45,13680 | 44,88160 | 90,01840 | 1,01111 | 1,111% | 41,84 |
| Xin | 44,67060 | 44,21585 | 88,88645 | 0,99840 | -0,160% | -6,03 |
| Xe | 44,64300 | 44,38600 | 89,02900 | 1,00000 | 0,000% | 0,00 |
| Xopt | 44,63620 | 44,38060 | 89,01680 | 0,99986 | -0,014% | -0,52 |

Na quarta coluna (m_{ag}) da tabela 6.3 acima, está apresentado o valor total da vazão de água de alimentação, que corresponde à soma dos valores das colunas x_3 e x_4 da tabela 6.3, que respectivamente são os valores da vazão de água de alimentação das colunas x_3 e x_4 da tabela 6.2. A quinta coluna (**Qsg norm**) apresenta a carga térmica do GV normalizada pela vazão de água de alimentação exata e pela entalpia, a qual se assume que é constante.

A coluna seis (**Qsg error**) corresponde à diferença entre a carga térmica do GV calculada e a carga térmica do GV exata expressa em percentual. A sétima coluna é a mesma medida de erro (**Error MWt**), mas expressa em megawatts térmicos (MW_t).

As linhas da tabela 6.3 correspondem a alguns testes típicos apresentados na tabela 6.2. A primeira linha corresponde ao teste R1, que representa um estimador com

as constantes a , b e c sem ajuste; A segunda linha (X_{in}) corresponde ao valor medido bruto, sem reconciliação de dados; a terceira linha (X_e) contém os valores exatos; a quarta linha (X_{opt}) apresenta os valores reconciliados obtidos com o método RDSMS.

Observando-se os resultados da tabela 6.3 verificou-se que ao se utilizar um estimador sem um ajuste adequado (R1), ocorreu uma propagação do erro de 1,111% no cálculo da carga térmica do GV, o que significa um erro de cerca de 41 MW térmicos considerando uma planta típica com produção de 3765 MWt. Este resultado (R1) indica problemas no fechamento do balanço de massa e energia e de acordo com a tabela 6.2, os erros grosseiros não foram detectados corretamente.

Adicionalmente, nota-se que devido à amplitude da propagação do erro no cálculo da carga térmica do GV, não é possível utilizar a margem de operação para aumentar a potência produzida dentro dos limites definidos pelo órgão regulador e pela análise de segurança do reator (102%).

Considerando-se a medida de água de alimentação sem reconciliação de dados (X_{in}), o erro de propagação indica que a medida da carga térmica do GV está abaixo do valor exato cerca de 6 MWt. Apesar do erro de propagação ser menor do que o exemplo anterior (R1), o mesmo ainda é significativo e com pouca margem de potência útil.

O resultado que considera a aplicação do método RDSMS proposto neste trabalho e indicado na linha X_{opt} da tabela 6.3, possui um erro de propagação cerca de 100 vezes menor do que o resultado utilizando o estimador robusto sem ajuste (R1) e cerca de 10 vezes menor do que o resultado que não utiliza a reconciliação de dados (X_{in}).

Nas tabelas 6.2 e 6.3 pode-se ver que a utilização do nosso método RDSMS, além de possuir uma propagação do erro menor, o número de erros grosseiros foi corretamente identificado e o valor estimado é preciso, fornecendo um valor muito acurado para o cálculo da carga térmica do GV e conseqüentemente para o cálculo da potência térmica do reator.

Nos exemplos das seções 6.2 e 6.3, o ajuste manual ou “off-line”, apresentou um ajuste próximo do ótimo, mas que pode levar à presença de um desvio na medida visto que os métodos de busca mencionados na literatura são métodos menos sofisticados,

como por exemplo, o método de busca binária, ou o método de busca Áurea (“Golden Search”). Dessa forma o método proposto mostrou-se também mais adequado à determinação das constantes de ajuste do M-estimador. Convém ressaltar que neste trabalho, não será realizada a comparação extensiva entre os dois tipos de método de ajuste das constantes do estimador.

Devido aos resultados promissores obtidos nos exemplos simulados apresentados nas seções 7.2 e 7.3, Na próxima seção será apresentado um exemplo com dados reais da aplicação do método RDSMS proposto aqui, quando aplicado ao balanço de massa em um circuito secundário em uma turbina a vapor, onde os dados reais foram obtidos da Usina Nuclear Angra 2, a qual tem a capacidade de gerar 1357 MWe ou 3765 MWt.

6.4 – Cálculo da Potência Térmica do Reator com Dados Reais obtidos na Usina Nuclear de Angra 2.

O exemplo apresentado nessa seção considera a aplicação do método simultâneo para a reconciliação de dados, identificação de erros grosseiros e seleção de modelos ao cálculo da potência térmica (desenvolvido nessa tese) de um reator nuclear do tipo PWR utilizando dados reais obtidos na Usina Nuclear de Angra 2.

A aplicação prática do método simultâneo apresentado neste trabalho é baseada nas orientações apresentadas na norma VDI-2048 e testadas no trabalho de VALDETARO e SCHIRRU (2011a), que mostra resultados efetivos em um cenário com dados simulados. Um resultado prático desse trabalho foi publicado em VALDETARO e SCHIRRU (2011b) e aqui serão considerados alguns aspectos adicionais relativos a aplicação do método e dos balanços de massa e energia.

Convém ressaltar mais uma vez, que os dados do processo relativos à Usina Nuclear de Angra 2 são utilizados para avaliar o desempenho do RDSMS em um contexto realístico e que essa pequena amostragem de dados não é própria para inferir qualquer condição particular de operação, desempenho ou segurança da Usina de Angra 2.

6.4.1 – Balanço de Massa Simplificado da Usina de Angra 2

A Usina de Angra 2 é uma usina nuclear do tipo PWR (Pressurized Water Reactor) com 4 loops e com potência nominal de 1357 MWe ou 3765 MWt. Na figura 6.3 está apresentada uma visão simplificada do circuito secundário da usina de Angra 2, com a indicação das medidas de vazão a serem utilizadas no balanço de massa da mesma.

Nessa figura as variáveis utilizadas são: a) m_{GV} , vazão total de vapor do Gerador de Vapor (soma das redundâncias 1 a 4); b) m_{GV1} a m_{GV4} , vazão de vapor individual de cada gerador de vapor; c) m_{ag} , vazão total de água de alimentação; d) m_{ag1} , vazão de água de alimentação da bomba de água de alimentação 1; e) m_{ag2} , vazão de água de alimentação da bomba de água de alimentação 2; f) m_{BD} , vazão da purga do GV; g) m_C , vazão de condensado; h) m_{A5} a m_{A7} , vazão das extrações A5 a A7; i) m_{HPC1} , vazão de retorno de condensado de alta pressão 1 e j) m_{HPC2} , vazão de retorno de condensado de alta pressão 2.

Seguindo o exemplo apresentado no capítulo 2 e baseado na norma VDI-2048 (2000), a vazão de vapor no ponto de entrada da turbina de alta pode ser determinada em função da vazão de vapor do GV, da vazão de água de alimentação e da vazão de condensado mais a vazão de água proveniente das extrações, conforme indicado nas equações (2.3) a (2.5), levando-se em conta a perda pela purga do GV e pela linha de desvio para o reaquecedor (linha em negrito).

Uma vez determinadas as equações relativas à vazão na entrada da turbina de alta é necessário estabelecer as relações entre as medidas que irá fornecer o balanço de massa do sistema, ou equações auxiliares ou restrições do processo, a fim de serem utilizadas no procedimento de reconciliação de dados, as quais estão indicadas abaixo:

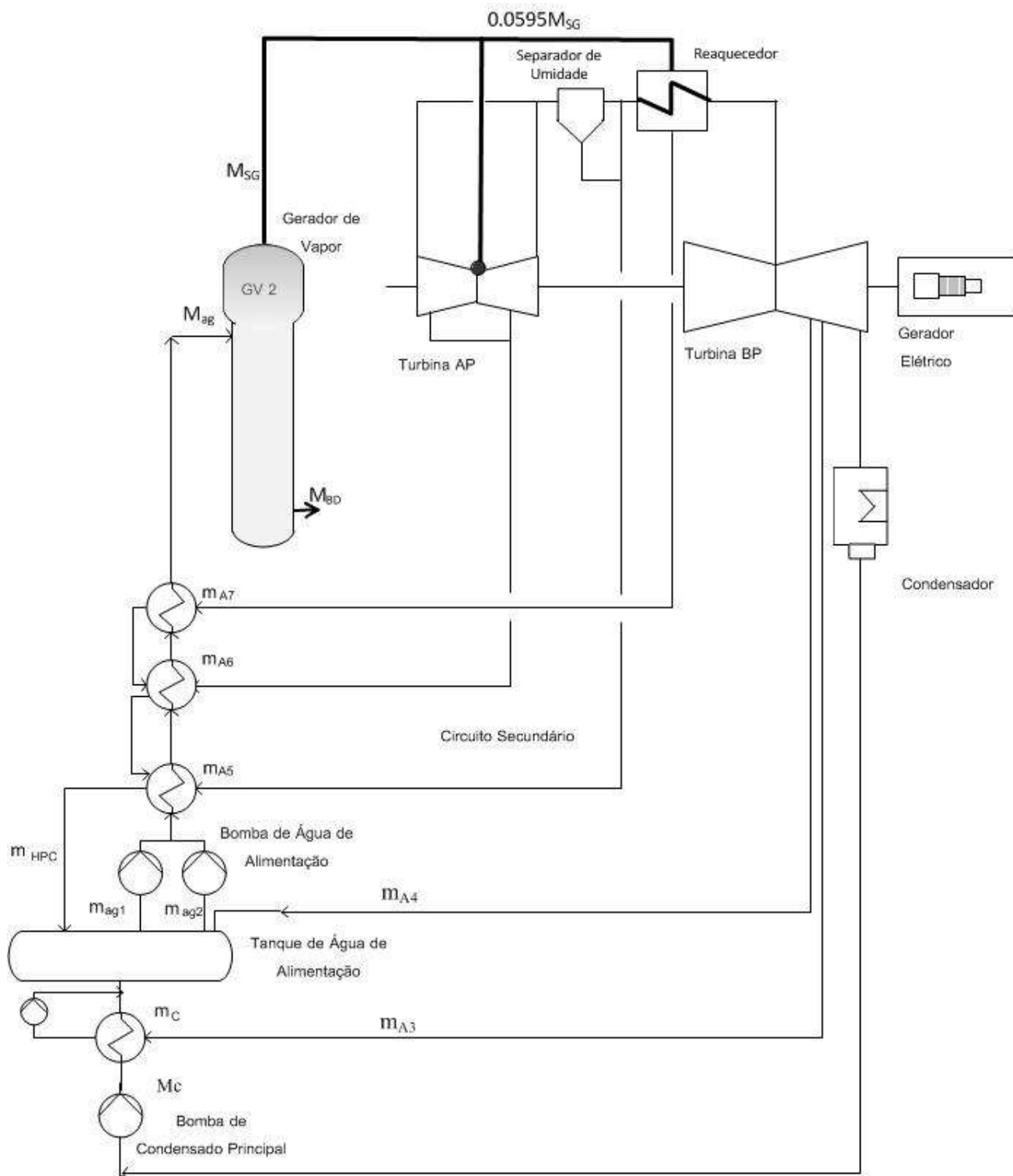


Figura 7.3 – Diagrama Simplificado do Circuito Secundário da Usina Nuclear de Angra 2.

$$M1 - M2 = 0 \quad (6.7)$$

$$M2 - M3 = 0 \quad (6.8)$$

$$m_{A7} + m_{A6} + m_{A5} - m_{HPC} = 0 \quad (6.9)$$

As equações (6.7) e (6.8) representam a diferença na vazão na entrada da turbina de alta, cujos valores devem ser iguais. Devem ser levados em conta nos balanços de massa e o desvio de vapor para o reaquecedor e a perda pela purga do GV. A equação 6.9 indica que o fluxo que entra no tanque de água de alimentação pela linha de retorno deve ser igual à soma dos fluxos de cada extração. Desta forma, as restrições do processo foram determinadas e o seguinte vetor de variáveis estimadas foi formado para ser usado na reconciliação de dados:

$$X = \begin{cases} [m_c, m_{HPC1}, m_{HPC2}, m_{ag1}, m_{ag2}, m_{GV1}, m_{GV2}, m_{GV3}, m_{GV4}, m_{A5a}, m_{A7a}, m_{A5b}, m_{A7b}, c] \\ ou \\ [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}] \end{cases} \quad (6.10)$$

Pode-se observar que no vetor x apresentado na equação (6.10) acima, não constam os valores medidos da extração A6 e os valores da purga do GV. O valor da vazão de água da extração A6 é estimado com base em medidas locais e estabelecido em torno de 136,00 Kg/s cada trem (2 trens). A vazão total dos 4 trens da purga do GV foi estabelecido em 10,8 Kg/s e foi estimada para efeito de simplificação da apresentação dos resultados, reduzindo o número de variáveis medidas e facilitando a apresentação.

A tabela 6.3 indica os resultados típicos de um ciclo da aplicação do método RDSMS ao cálculo da reconciliação de dados do balanço de massa simplificado da Usina de Angra 2, onde em cada ciclo, neste teste, foram realizadas 120 iterações. O método RDSMS foi apresentado no capítulo 5 e consiste em minimizar o índice AICR utilizando o estimador de três partes de Hampel sujeito as restrições do processo,

conforme apresentado na equação (5.7). Para reduzir o número de colunas da referida tabela e permitir melhor apresentação, a coluna m_{HPC} indica a soma das vazões de condensado de alta pressão 1 e 2 e as colunas m_{A5} , m_{A6} e m_{A7} apresentam a soma dos dois trens das referidas extrações.

Tabela 6.3 - Resultados do cálculo do balanço de massa simplificado da Usina de Angra 2 pelo método RDSMS.

| Iterações (ni) | m_c | m_{HPC} | m_{ag1} | m_{ag2} | m_{GV1} | m_{GV2} | m_{GV3} | m_{GV4} | m_{a5} | m_{a6} | m_{a7} | c | MWt | % |
|----------------|------------------|-----------------|------------------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------------|-----------------|---------------|---------------|------------|
| 4 | 1315,3428 | 647,8471 | 1073,3778 | 1047,6288 | 536,0469 | 521,5322 | 525,7899 | 531,2549 | 122,7008 | 273,14 | 243,3715 | 3,8477 | 3887,4 | 3,3 |
| 10 | 1326,7450 | 627,2094 | 1055,9562 | 1050,6395 | 530,2581 | 518,6706 | 526,6654 | 524,9142 | 121,1546 | 273,14 | 226,3462 | 1,5618 | 3861,5 | 2,6 |
| 20 | 1326,7450 | 627,2094 | 1055,9562 | 1050,6395 | 530,2581 | 518,6706 | 526,6654 | 524,9142 | 121,1546 | 273,14 | 226,3462 | 1,5618 | 3867,9 | 2,7 |
| 30 | 1320,3547 | 627,1570 | 1050,3222 | 1048,2052 | 525,9407 | 523,8441 | 528,4446 | 525,7760 | 121,1833 | 273,14 | 230,7229 | 1,3153 | 3867,9 | 2,7 |
| 40 | 1321,7372 | 618,0936 | 1047,2848 | 1044,3642 | 523,0895 | 524,6639 | 526,4466 | 522,5857 | 121,1062 | 273,14 | 225,2253 | 1,4677 | 3846,0 | 2,2 |
| 50 | 1320,1388 | 618,5314 | 1045,7909 | 1046,2593 | 521,6799 | 523,0364 | 525,4445 | 521,9169 | 121,3729 | 273,14 | 225,3298 | 1,4252 | 3846,0 | 2,2 |
| 60 | 1320,2849 | 618,9812 | 1045,9044 | 1045,5387 | 521,5092 | 523,5445 | 525,3578 | 521,9002 | 121,4289 | 273,14 | 225,6941 | 1,4264 | 3847,0 | 2,2 |
| 70 | 1320,4403 | 618,6951 | 1045,7600 | 1045,3127 | 521,5294 | 523,3641 | 525,5607 | 522,1744 | 121,5773 | 273,14 | 225,3262 | 1,4538 | 3844,5 | 2,1 |
| 80 | 1318,6936 | 619,1317 | 1044,9984 | 1045,2073 | 521,5055 | 522,7315 | 525,8896 | 521,1120 | 121,3818 | 273,14 | 225,3597 | 1,4176 | 3844,2 | 2,1 |
| 90 | 1318,4264 | 618,7742 | 1044,9797 | 1044,8017 | 521,5131 | 522,5213 | 525,9200 | 521,1616 | 121,4696 | 273,14 | 225,3228 | 1,4191 | 3844,5 | 2,1 |
| 100 | 1318,7207 | 619,0038 | 1045,0504 | 1045,2079 | 521,4335 | 522,5636 | 525,8687 | 521,1502 | 121,3671 | 273,14 | 225,4870 | 1,4202 | 3844,0 | 2,1 |
| 110 | 1318,6390 | 618,7871 | 1044,6498 | 1045,1967 | 521,4194 | 522,5343 | 525,8682 | 521,0568 | 121,3694 | 273,14 | 225,5050 | 1,4236 | 3843,8 | 2,1 |
| 120 | 1318,6334 | 619,0327 | 1044,8015 | 1045,3147 | 521,2492 | 522,6836 | 525,8619 | 520,9980 | 121,4078 | 273,14 | 225,5808 | 1,4225 | 3843,6 | 2,1 |
| Validado | 1321,320 | 621,84 | 1064,780 | 1034,190 | 520,442 | 512,426 | 527,350 | 509,677 | 121,2227 | --- | 226,6400 | --- | 3766,3 | 0,03 |

Pode-se verificar na tabela 6.3 acima, que os resultados obtidos na iteração N^o 120 (em negrito) estão muito próximos ao valor de referência (linha **validado** na referida tabela). Os valores de referência foram obtidos por meio da aplicação da reconciliação de dados clássica, calculadas pelo software implementado em Angra 2 (Software VALI, da empresa BELSIM) para monitoração de performance da Usina de Angra 2.

Neste exemplo, foram utilizadas cerca de 200 medidas adquiridas em intervalos de 1 minuto. Não foram incluídos erros grosseiros. Em cada ciclo (um passo de deslocamento da janela de tempo) foram utilizadas 280 partículas (np) e 120 iterações (ni) para concluir o ajuste simultâneo a cada ciclo. O horizonte de medidas (H) é de 100 medidas. Após a iteração 50, os valores globais obtidos no PSO, ou seja, o vetor solução dado pela equação (6.10), atingiu a região de viabilidade, onde os valores estimados satisfazem os balanço de massa.

Para calcular a carga térmica do gerador de vapor, foram utilizados os seguintes valores de entalpia de entrada e saída do GV estabelecidos na condição de operação da usina durante a aquisição dos dados: $h_e=935.55$ KJ/Kg $h_s= 2773.8$ KJ/Kg, respectivamente.

O resultado final do cálculo da potência térmica da usina de Angra 2 pelo método RDSMS indicado na coluna MW_t na tabela 6.3 está indicando um valor de 3843,6 MW_t em um patamar próximo do valor calculado pelo software implementado em Angra 2 para monitoração do desempenho térmico da Usina de Angra 2, que foi de 3766,3 MW_t. Embora o balanço de energia não tenha sido considerado, os resultados são coerentes com os obtidos pelo sistema da Usina de Angra 2 (software VALI). A diferença entre o valor calculado da potência do reator pelo método RDSMS e a potência do reator validada pelo software VALI foi de 77,3 MW_t o que corresponde a 2,1 % da potência do reator validada. Isto representa um valor muito próximo da referência, considerando que o método RDSMS foi aplicado em um balanço de massa simplificado, mostrando que o método RDSMS é efetivo, ainda que com simplificações no modelo.

Aplicando uma aproximação grosseira, pode-se estimar a potência do reator considerando o balanço de energia, descontando-se a potência consumida pelas Bombas de Refrigeração do Reator e pela purga do GV, que é de aproximadamente 40 MW_t, do valor obtido na tabela 7.3 de 3843,6 MW_t, resultando em cerca de 3803,6 MW_t, o que corresponde a 0,99 % da potência do reator validada. Esse valor relativo à potência térmica do reator pode ser ainda melhorado ao se considerar um balanço térmico mais detalhado.

O teste foi realizado a partir de dados obtidos do processo gravados em arquivo e esses dados foram lidos pelo programa de ajuste automático, que foi implementado no programa MATLAB e executado em um microcomputador Core 2 Duo com 1.6 GHz.

CAPÍTULO 7:

CONCLUSÕES

7.1 - Introdução

Nesse capítulo serão apresentadas as conclusões gerais obtidas nesta tese, que resultaram do desenvolvimento do método para Reconciliação Robusta de Dados e Seleção de Modelo Simultânea (RDSMS). O método proposto é voltado para a aplicação na monitoração “on-line” e aqui, em particular, o mesmo está voltado para o cálculo da potência térmica de um reator nuclear do tipo PWR, incluindo exemplos simulados e uma aplicação prática com dados reais obtidos da Usina de Angra 2.

Na seção relativa às conclusões, serão apresentados ainda os comentários que contribuíram para a originalidade desta tese de doutorado e na última seção serão apresentadas sugestões para trabalhos futuros.

7.2 – Conclusões Gerais

Neste trabalho foi apresentada uma visão geral dos principais métodos para Reconciliação de Dados e aqueles que fundamentam o desenvolvimento do método RDSMS. De forma a embasar o desenvolvimento da referida metodologia e reforçar a importância da aplicação da estratégia de Reconciliação de Dados e Identificação de Erros Grosseiros, foram abordados ainda o desenvolvimento do cálculo da potência térmica de um reator nuclear tipo PWR utilizando um sistema simplificado do circuito secundário de uma usina nuclear do tipo PWR, cujo modelo foi baseado na norma VDI-2048 (2000). Também foi apresentada uma visão geral sobre aspectos da estatística robusta, de estimadores robustos e sobre o algoritmo de otimização utilizado no desenvolvimento deste trabalho, que é o algoritmo de otimização baseado em enxame de partículas (PSO) padrão.

O método desenvolvido nesse trabalho é baseado na minimização direta do Critério de Informação de Akaike Robusto (AICR), que é próprio para a utilização com estimadores robustos. O ajuste das constantes do estimador robusto é incorporado ao problema de minimização como mais um objetivo, o que permite resolver simultaneamente o problema de reconciliação de dados e a seleção de modelo de probabilidade. Desta forma, é eliminado o ajuste em duas fases, ou seja, não há a necessidade de ajustar primeiro as constantes do estimador e num segundo passo resolver o problema da reconciliação de dados, o que é considerado uma melhoria do método proposto RDSMS em relação a outros métodos de RD.

Com a utilização do Critério de Informação de Akaike Robusto não é mais necessário assumir que a contaminação do erro possui uma distribuição definida, usualmente a distribuição Normal, o que permitiu eliminar os problemas que decorrem dessa hipótese.

Considerando que outros métodos de reconciliação de dados acabam por utilizar simulações para calcular índices de desempenho, como o cálculo da eficiência relativa, a qual é determinada por meio de simulações e uso do método Monte Carlo, o método RDSMS também apresenta vantagens no uso desses métodos (seção 5.1). Nele o cálculo é feito de forma exata, determinando o conjunto principal de dados e minimizando os índices de desempenho do estimador robusto, como a sensibilidade a erros grosseiros, e diminuindo o desvio assintótico (“bias”). Assim, o uso do índice AICR proposto por RONCHETTI (1985, 1997a) nesta metodologia junto com o PSO forneceu um caminho ou sistemática geral para a reconciliação robusta de dados, identificação de erros grosseiros, mesmo considerando problemas não lineares, com diversas variáveis ou restrições.

O desempenho do método proposto apresentou resultados positivos e efetivos, tanto em exemplos com dados simulados (PAI e FISHER, 1988 e VDI-2048, 2000) como em um exemplo de cunho prático com dados reais da Usina de Angra 2 (VALDETARO e SCHIRRU, 2011b). Esses resultados estão apresentados nos itens 6.2, 6.3 e 6.4, respectivamente.

Pode-se observar nos exemplos, que o estimador de três partes de Hampel apresentou robustez quando na presença de erros grosseiros, além de apresentar

resultados não tendenciosos e com a capacidade de identificar corretamente os erros sistemáticos incorporados (“sensor travado”).

Outras vantagens são a facilidade de implementação do método RDSMS e o uso de um algoritmo de otimização global eficiente, que é o PSO, implementado de forma padrão. Como a função objetivo depende de um horizonte de medidas para realizar a estimação robusta dos dados, o método RDSMS pode ser implementado de forma dinâmica utilizando uma janela de dados de tamanho h , que se desloca a cada iteração. Dessa forma, se a característica da estatística dos dados sofrer alguma alteração, o método RDSMS se auto-ajusta, calculando as constantes do estimador a cada iteração.

Os resultados dos exemplos com dados simulados e da aplicação prática com dados reais da Usina de Angra 2 se mostraram efetivos. No exemplo do item 6.1 com o sistema não linear proposto por PAI e FISHER (1988), todos os erros grosseiros foram identificados corretamente nas diversas componentes do vetor de resultados e o valor ótimo obtido com o método RDSMS (X_{opt}) apresentou uma precisão no mínimo de uma casa decimal e até a segunda casa decimal no máximo. Os resultados podem ser melhorados escolhendo maior número de partículas ou iterações.

No segundo exemplo apresentado no item 6.2, relativo ao balanço de massa simplificado baseado na norma VDI-2048, também apresentou resultado semelhante ao caso anterior. Todos os erros grosseiros foram identificados corretamente, na componente correta do vetor solução, e o resultado ótimo (X_{opt}) se aproximou do resultado exato (X_e) com no mínimo uma casa decimal e no máximo duas casas decimais. Assim o método proposto aqui (RDSMS) mostrou evidências de sua eficiência na realização da reconciliação robusta de dados e na identificação simultânea de erros grosseiros.

Para avaliação do método RDSMS em um cenário mais realista, ele foi testado em um balanço de massa simplificado do circuito secundário da usina nuclear de Angra 2 com dados reais. Os resultados apontaram para uma reconciliação de dados eficiente, onde a potência do reator foi estimada com cerca de 2,1% de diferença entre o valor de referência (Linha Validado na tabela 6.3) e o valor calculado obtido na iteração nº 120 da referida tabela. Nesse caso não foram introduzidos erros grosseiros, pois o objetivo foi avaliar o desempenho do método RDSMS nas condições normais de operação. Ao se

considerar uma estimativa de balanço de energia, mesmo de forma aproximada, a diferença entre o valor estimado e o valor validado foi de 0,9%, o que é um fator que indica a possibilidade de aplicação desse método em uma aplicação em um cenário real.

Uma execução de uma iteração completa dura entre 3 a 4 minutos para completar e esse tempo depende do número de partículas e do número de iterações. Em cenários reais, um período de aquisição utilizado em diversos softwares comerciais para aplicar a RD e IEG é de cerca 15 minutos, o que mostra o potencial do método para a monitoração on-line. Deve-se ressaltar que em aplicações reais o número de variáveis é muito grande e uma comparação direta deve ser feita somente após testes extensivos de avaliação do método proposto.

O método de Reconciliação de Dados Robusta com Seleção de Modelo Simultânea apresentou características inovadoras e a principal é a possibilidade de efetuar a sintonia do estimador robusto de forma simultânea ao problema de reconciliação de dados e identificação de erros grosseiros, além de outras vantagens mencionadas acima. Entretanto, outros trabalhos devem ser desenvolvidos para permitir uma avaliação mais extensa e para se ter uma visão mais detalhada do comportamento do referido método.

7.3 – Sugestões para Trabalhos Futuros

Para trabalhos futuros, pode-se explorar a utilização de outros estimadores robustos e a avaliação de desempenho dos mesmos quando aplicado no método proposto.

Outra questão em aberto é a relação entre as constantes de ajuste do estimador de Hampel e a proporcionalidade entre elas, que pode ser explorada a fim de se obter proporções que atendam melhores índices de robustez, como a sensibilidade a desvios na medida (λ).

Outra linha de trabalho pode ser a avaliação extensiva do desempenho do método proposto, considerando-se um melhor detalhamento nos balanços de massa e de energia e diversas condições de operação em determinado processo, sendo que em

condições cujo problema exija um número muito grande de variáveis, avaliar o comportamento do algoritmo PSO e a possibilidade de uso de outros algoritmos de otimização.

CAPÍTULO 8:

REFERÊNCIAS BIBLIOGRÁFICAS

- AGOSTINELLI, C., 2002. Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*. 56, 289-300.
- AKAIKE, H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control*. 19, 716-723.
- ALBUQUERQUE, J.S., BIEGLER, L.T., 1996. Data reconciliation and gross error detection for dynamic system. *AIChE Journal*. 42(10), 2841-2856.
- ANDRADE, L. A., MARTINEZ, C.B., FILHO, J.N., AGUIRRE, L. A., 2002. Estudo comparativo dos métodos de medição de vazão – Uma aplicação em comissionamento de turbinas hidráulicas, In: CPH Internal Report, available at www.cph.eng.ufmg.br/docscph/matevento15.pdf, accessed in 08/Jan/2009 18:30.
- ARORA, N., BIEGLER, L.T., 2001. Redescending estimators for data reconciliation and parameter estimation. *Computers and Chemical Engineering*. 25, 1585-1599.
- AZOLA E., SAMPAIO M.S., VALDETARO E.D., 2009, Utilização da reconciliação de dados para implantação de um programa de aumento de desempenho da unidade 2 da central nuclear almirante Álvaro Alberto, In: XX SNPTEE, Seminário Nacional de Produção e Transmissão de Energia Elétrica, Nov, Rio de Janeiro, Brasil.
- BASU, A., PALIWAL, K.K., 1989, Robust M-estimates and generalized M-estimates for autoregressive parameter estimation, In: Proceedings of TENCON 89, Fourth IEEE Region 10 International Conference, Bombay, 355-358.
- BOZDOGAN, H., 2000, Akaike's Information Criterion and Recent Developments in Information Complexity *Journal of Mathematical Psychology* 44, 62-91.

- BRATON, D., KENNEDY J., 2007. Defining a standard for particle swarm optimization, In: Proceedings of the 2007 IEEE Swarm Intelligence Symposium (SIS 2007).
- CROWE, C.M., CAMPOS, Y.A.G., HRYMAK, A, 1983. Reconciliation of process flow rates by matrix projection I: Linear case. *AIChE Journal*. 29, 881-888.
- FELDMAN, R. N. (2007), “Reconciliação de dados em tempo real para monitoração e detecção de falhas em terminal de transporte e armazenamento de derivados de petróleo”, COPPE/UFRJ, M.Sc. Thesis, Chemical Engineering Department.
- GRAUF, E., JANSKY, J., LANGENSTEIN, M., 2000. Reconciliation of process data in Nuclear Power Plants (NPPs), In: Proceedings of ICONE 8 , 8th International Conference on Nuclear Engineering, April 2-6, Baltimore, MD USA.
- HAMPEL, F.R., 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 69. No 346., 383-393.
- HAMPEL, F.R., RONCHETTI, E.M, ROUSSEEUW, P.J., STAHEL, W.A., 1986. *Robust Statistics - The Approach based on Influence Functions*, John Wiley & Sons, Wiley Series in Probability and Mathematical Statistics.
- HUBER, P.J., 1964. Robust Estimation of a Location Parameter, In: *The Annals of Mathematical Statistics*, 35 (1), 73-101.
- HUBER, P.J., 1981. *Robust Statistics*. John Wiley & Sons, Wiley Series in Probability and Mathematical Statistics.
- JANSKY, A., 2006. Financial benefits of process data reconciliation in power generating plants, In: Proceedings of ICONE14, 14th International Conference on Nuclear Engineering, July 17-20, Miami, USA.
- JANSKY, A., 2007. Increasing plant efficiency and safety with online process data reconciliation, In: Proceedings of ICONE15, 15th International Conference on Nuclear Engineering, April 22-26, Nagoya, Japan.

- JOHNSTON, L.P.M., KRAMER, M.A., 1995. Maximum likelihood data rectification: Steady-state systems. *AIChE Journal*. 41(11), 2415–2426.
- KENNEDY, J., EBERHART, R.C., 1995. Particle Swarm Optimization, In: Proceedings on feedback mechanism, IEEE International Conf. on Neural Networks. VI, 1942-1948.
- KENNEDY, J., 2007. Some issues and practices for particle swarms, In: Proceedings of the 2007 IEEE Swarm Intelligence Symposium (SIS 2007).
- KUEHN, D.R., DAVIDSON, H., 1961. Computer control. II. Mathematics for control. *Chemical Engineering Progress*. 57, 44-47.
- MEI, C., SU, H., CHU, J., 2007. Detection of gross errors using mixed integer optimization approach in process industry, In: *Journal of Zhejiang University SCIENCE A*. 8(6), 904-909.
- MORAD K., YOUNG B.R., SVRCEK W.Y., 2005. Rectification of plant measurements using a statistical framework. *Computers and Chemical Engineering*. 29, 919-940.
- MOROS, R., KALIES, H., REX, H.G., & SCHAFFARCZYK, S., 1996. A genetic algorithm for generating initial parameter estimations for kinetic models of catalytic processes. *Computers and Chemical Engineering*. 20(10), 1257-1270.
- OZYURT, D.B., PIKE, R.W., 2004. Theory and practice of simultaneous data reconciliation and gross error detection for chemical process. *Computers and Chemical Engineering*. 28, 381-402.
- PAI, C.C.D., FISHER, G.D., 1988. Application of Broyden's method to reconciliation of nonlinearly constrained data. *AIChE Journal*. 34(5), 873-876.
- PRAKOTPOL, D., SRINOPHAKUN, T., 2003. GAPinch: Genetic algorithm toolbox for water pinch technology. *Chemical Engineering and Processing*. 43(28), 203-217.
- PRATA, D. M., 2009, *Reconciliação Robusta de Dados para Monitoramento em Tempo Real/ Diego Martinez Prata. Tese de Doutorado– Rio de Janeiro: UFRJ/COPPE, jun/2009.*

- PRATA, D.M., PINTO, J.C., LIMA, E.L., 2008. Comparative analysis of robust estimators on nonlinear dynamic data reconciliation. *Computer-aided Chemical Engineering*. 25, 501–506.
- PRATA, D.M., SCHWAAB, M., LIMA, E.L., PINTO, J.C., 2009. Nonlinear dynamic data reconciliation and parameter estimation through particle swarm optimization: Application for an industrial polypropylene reactor. *Chemical Engineering Science*. 64, 3953-3967.
- PRATA, D.M., PINTO, J.C., LIMA, E.L., 2010. Simultaneous robust data reconciliation and gross error detection through particle swarm optimization for an industrial polypropylene reactor. *Chemical Engineering Science*. 65, 4943-4954.
- RONCHETTI, E., 1985. Robust model selection in regression. *Statist. Probab. Lett.* 3, 21-23.
- RONCHETTI, E., 1997a. Robustness aspects of model choice. *Statistica Sinica*. 7, 327-338.
- RONCHETTI, E., 1997b. Robust inference by influence functions, *Journal of Statistical Planning and Inference*. 57, 59-72.
- ROUSSEEUW. P.J., LEROY, A.M, 1987. *Robust Regression and Outlier Detection*, John Wiley & Sons, Wiley Series in Probability and Mathematical Statistics.
- ROUSSEEUW P.J., CROUX, C., 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*. 88. No. 424, 1273-1283.
- SERTH, R.W., HEENAN, W.A., 1986. Gross error detection and data reconciliation in steam-metering systems. *AIChE Journal*. 32(5), 733-742.
- SHI, Y., EBERHART, R.C., 1998. A modified Particle Swarm Optimizer, In: *Proceedings of the 1998 IEEE Congress on Evolutionary Computation*, Anchorage, AK.

- SODERSTROM, T.A., HIMMELBLAU, D.M., EDGAR, T.F., 2000. A mixed integer optimization approach for simultaneous data reconciliation and identification of measurement bias. *Control Engineering Practice*. **9**, 869-876.
- SPANOS, A., 2010. Akaike-type Criteria and the Reliability of Inference: Model Selection versus Statistical Model Specification. *Journal of Econometrics*, **158**, 204-220.
- STREIT, S., LANGENSTEIN, M., LAIPPLE, B., EITSCHBERGER, H., 2005. A new method for evaluation and correction of thermal reactor power and present operational applications, In: *Proceedings of ICONE13, 13th. International Conference on Nuclear Engineering*, May 16-20, Beijing, China.
- TJOA, I.B., BIEGLER, L.T., 1991. Simultaneous strategy for data reconciliation and gross error detection of nonlinear systems. *Computers and Chemical Engineering*. **15**(10), 679-690.
- VALDETARO, E.D., SCHIRRU, R., 2009. Particle Swarm Optimization applied to data reconciliation in nuclear power plant, In: *International Nuclear Atlantic Conference – INAC2009*, Sep. 27 to Oct.2, Rio de Janeiro, RJ, Brasil.
- VALDETARO, E.D., SCHIRRU, R., 2011a. Simultaneous Model Selection, Data Reconciliation, and Outlier Detection with Swarm Intelligence in a Thermal Reactor Power Calculation. *Annals of Nuclear Energy*. **38**, 1820-1832l.
- VALDETARO, E.D., SCHIRRU, R., 2011b. Robust Data Reconciliation and Outlier Detection with Swarm Intelligence in a Thermal Reactor Power Calculation In: *International Nuclear Atlantic Conference – INAC2011*, Out. 24 to Out.28, Rio de Janeiro, RJ, Brasil.
- VDI 2048 Part I, 2000. Uncertainties of measurement during acceptance tests on energy-conversion and power plants- fundamentals.
- WASANAPRADIT, T., 2000. Solving nonlinear mixed integer programming using genetic algorithm”, Master Thesis, King Mongkut University of Technology Thonburi, Bangkok, Thailand. Available: fengtcs@ku.ac.th.

WONGRAT, W., SRINOPHAKUN, T., SRINOPHAKUN, P., 2005. Modified genetic algorithm for nonlinear data reconciliation. *Computers and Chemical Engineering*. 29, 1059-1067.

YAMAMURA, K., NAKAJIMA, M., MATSUYAMA, H., 1988. Detection of gross errors in process data using mass and energy balances. *International Chemical Engineering*. 28(1), 91-96.

ZHOU, L., SU, H., CHU J., 2006. A new method to solve robust data reconciliation in nonlinear process, *Chinese J. Chem. Eng.* 14(3), 357-363.
