UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

Avaliação de Estratégias para Reconciliação de Dados e Detecção de Erros Grosseiros

DISSERTAÇÃO DE MESTRADO

ANDREA CABRAL FARIAS

Porto Alegre
2009

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

Avaliação de Estratégias para Reconciliação de Dados e Detecção de Erros Grosseiros

Andrea Cabral Farias

Dissertação de Mestrado apresentada como requisito parcial para obtenção do título de Mestre em Engenharia

Área de concentração:

Orientador:
Prof. Dr. Argimiro Resende Secchi

Porto Alegre

2009

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Dissertação *Avaliação de Estratégias para Reconciliação de Dados e Detecção de Erros Grosseiros*, elaborada por Andrea Cabral Farias, como requisito parcial para obtenção do Grau de Mestre em Engenharia.

Comissão Examinadora:

Dr. Oscar Rotava

Prof. Dr. Luís Gustavo Soares Longhi

Prof. Dr. Marcelo Farenzena

Agradecimentos

Em primeiro lugar eu gostaria de agradecer à minha família. À minha mãe pelo exemplo, pelo estímulo para continuar *"não importa o que aconteça"*, pelo carinho, apoio e dedicação. Ao meu irmão pela amizade, parceria, atenção e por ter dedicado tanto do seu tempo - praticamente meu co-orientador neste trabalho. À minha avó pelo apoio incondicional, mesmo não entendendo *uma vírgula* do que eu estava fazendo.

Gostaria de agradecer ao meu namorado Leandro pela sua paciência (e às vezes pela falta dela!), pelo seu apoio e por sua parceria, assim como da sua (nossa) família e seus (nossos) amigos.

Gostaria de agradecer ao amigo Dr. Ricardo G. Duraiski, pela paciência e boa vontade em tentar me ensinar a programar. Ao Professor e amigo ("e primo") Marcelo Farenzena pelos papos, discussões, dicas e divagações. Ao pessoal do GIMSCOP, especialmente à Débora - minha sempre amiga, desde a graduação. À Gabriela, ao Escobar e à Bruna pelas risadas e papos. Aos meus colegas: Pedro I., João H., Fernanda B., Marcelo O., Cecília, Giovana, Thais, Fábio, Guga, Giovani, Carine e Dudu – que me deram "uma baita mão" no início, enquanto eu trabalhava no pólo petroquímico 24 horas por dia e cursava as disciplinas obrigatórias. Também não poderia de deixar de agradecer à minha grande amiga Fernanda Vargas, que foi a pessoa mais presente nos últimos 10 anos e sempre esteve do meu lado me apoiando nos momentos bons e ruins.

Gostaria de agradecer ao meu orientador Argimiro e a sua esposa Sirley, que sempre foram um exemplo pra mim e por quem eu tenho imensa admiração e gratidão. Obrigada pela confiança e pelo seu carinho, atenção, paciência, palavras amigas nos poucos momentos de frustração e amizade.

Gostaria de agradecer ao DEQUI e à UFRGS.

Resumo

O sistema de reconciliação de dados trata de um problema advindo da evolução das técnicas de medição, aquisição e armazenamento de dados. Este tem o papel de garantir a consistência destes dados, utilizando a redundância das variáveis medidas e um modelo estatístico da medição para aumentar a precisão dos dados. O procedimento completo tem por objetivo que as equações de conservação sejam satisfeitas, tratando dos erros aleatórios inerentes ao processo de medição e que eventuais erros grosseiros sejam detectados e corrigidos. Estas duas últimas atribuições referem-se aos dois problemas tratados neste trabalho: avaliação de técnicas de reconciliação de dados e para a detecção de erros grosseiros. O objetivo deste trabalho é comparar diferentes técnicas por meio de um estudo completo de reconciliação de dados e detecção de erros grosseiros. Este foi baseado em simulações determinísticas e simulações Monte Carlo para verificar o desempenho das estratégias frente aos parâmetros que influenciam cada etapa do procedimento. Em reconciliação de dados foi avaliada a influência da topologia e do pré-tratamento de dados na qualidade final da estimação. Já para etapa de detecção de erros grosseiros foram avaliadas sete estratégias diferentes, realizando uma comparação entre as mesmas com base em um estudo combinatorial completo. Avaliou-se a influência da topologia e foram levantadas as curvas de poder de detecção. Com base nestes resultados escolheu-se um critério para que os algoritmos fossem sintonizados de maneira que a comparação entre eles fosse justa. Após a sintonia avaliou-se a utilização do pré-tratamento de dados. Além das estratégias de detecção tradicionais utilizaram-se também técnicas de reconciliação robusta. O desempenho destas foi comparado com os resultados obtidos nas etapas anteriores. Como conseqüência deste estudo completo, foi proposta uma nova estratégia de detecção de erros grosseiros, baseada em estatística robusta. O seu desenvolvimento foi demonstrado e a validação foi realizada por comparação com os resultados obtidos neste trabalho e com um caso reportado na literatura.

Abstract

The data reconciliation system deal with a problem originated from the evolution of measurement techniques, data acquisition, and data storage. This system plays the role of guaranteeing the data consistency; it uses the measured variables redundancy and a statistical measurement model to improve accuracy. The main goal of the complete procedure is to satisfy the conservation equations, treating the random errors inherent of the measurement process and detecting eventual gross errors. The last two attributes are the issues of this work: the evaluation of data reconciliation techniques and the problem of gross error detection. The goal of this work is to compare different techniques by a complete data reconciliation and gross error detection study, based on deterministic and Monte Carlo simulations to verify the performances of the strategies as functions of the parameters that influence each step of the procedure. In data reconciliation, the influence of the topology and data pre-treatment in the quality of the estimates were investigated. Furthermore, dealing with the gross error detection step, seven different strategies were compared by means of a complete combinatorial study. The influence of topology was studied and the power curves were obtained. Based on these results, a criterion to tune the algorithms was chosen in the manner of guaranteeing a fair comparison between them. After the tuning step, the use of data pre-treatment was investigated. To complete the study, robust data reconciliation techniques were also used, and their performances were compared with the results attained in the precedent sections. As a product of this study, a new gross error detection strategy was proposed, based on robust statistics. The development steps were showed and the new method was validated based on comparison with the results obtained in this work and with a case study from the literature.

Sumário

Lista de figuras

Lista de tabelas

Lista de abreviaturas

AVTI	Média de erros tipo I
CN	Função normal contaminada
DDR	“Dynamic data reconciliation”
DEG	Detecção de erros grosseiros
EG	Erro grosseiro
EKF	“Extended Kalman filter”
ERest	Erro da estimação
FIR	Filtro de resposta à impulso finito
GLR	Teste da razão da máxima verossimilhança generalizada
GT	Teste global
ICS	“Imbalance correlation strategy”
IF	Função de Influência
IIR	Filtro de resposta à impulso infinito
IMT	Teste iterativo da medição
IMTr1	Versão eliminatória, iterativa e robusta do IMT
IMTr2	Versão eliminatória, simultanea e robusta do IMT
LCT	Técnica da combinação linear
MCGLR	GLR com restrições modificado
MHE	“Moving horizon estimator”
MILP	Programação linear inteira mista
MIMT	Teste iterativo da medição modificado
MINLP	Programação não linear inteira mista
MLE	Estimação de máxima verossimilhança
MQP	Mínimos quadrados ponderados
MSCS	Estatégia de compensação seriada modificada
MT	Teste da medição
MTMP	Teste da medição de máxima potencia
MUBET	“Modified unbiased estimation technique”
NDDR	“Nonlinear dynamic data reconciliation”
NT	Teste nodal
NTMT	Teste nodal de máxima potencia
OP	Poder de detecção global
OPF	Fração esperada de perfeita identificação
PC	Componente principal
PCA	Análise dos componentes principais
QP	Programação quadrática
RD	Reconciliação de dados
SDCD	Sistema de controle digital distribuído
SEGE	Estimação simultânea de erros grosseiros
SPRT	Teste sequencial da razão de probabilidade
SQP	Programação quadrática seqüencial

SSCS	Estratégia de compensação seriada simples
SVG	Filtro de Savitzky-Golay
TEEq	Teoria dos erros equivalentes
TER	Redução do Erro Total
UBET	“Unbiased estimation technique”

Lista de símbolos

e	erro total da medição
ε	erro aleatório
δ	erro grosseiro (magnitude)
y	valor medido
x	valor verdadeiro
W	matriz de variância-covariância da medição
σ	desvio padrão
χ^2	distribuição qui quadrado
ν	número de graus de liberdade
Γ	função gama
A	matriz de incidência
aj	ajustabilidade
dt	detectabilidade
^	refere-se à variável estimada ou reconciliada
h	conjunto de restrições de igualdade
g	conjunto de restrições de desigualdade
p	parâmetros
m	número de variáveis medidas
n	número de equações de restrição
λ	multiplicadores de Lagrange
μ	média ou valor esperado de x
ρ	função objetivo de Huber
γ	estatística usada no GT
a	ajuste calculado na reconciliação
Za	estatística usada no MT
Zr	estatística usada no NT
α	nível de confiança
α^*	nível de confiança modificado
r	resíduo das restrições (A.x)
H_0	hipótese nula
H_1	hipótese alternativa
d	ajuste calculado na reconciliação ponderado por W
pr	vetor das componentes principais
Λ	valores característicos de W
Ur	matriz de vetores característicos ortonormalizados de V
k	número de PC retidas
w	variável aleatória

Capítulo 1

Introdução

1.1 Reconciliação de Dados e Detecção de erros grosseiros

Para monitorar efetivamente uma planta industrial, controlar a produção e garantir sua operação é necessário que se saiba o estado real da planta em qualquer momento desejado. Com este propósito, um grande número de variáveis de processo são medidas e seus valores armazenados em banco de dados em tempo real. Com o avanço das técnicas de medição e modelagem, atreladas ao avanço das técnicas computacionais, a quantidade de dados de processo armazenada só tende a aumentar (Kongsjahju, Rollins e Bascuñana, 2000).

Segundo Bascur e Linares (2006), um dos maiores desafios para análise de desempenho operacional é a aquisição, validação e reconciliação da informação do processo. Informações confiáveis e validadas sobre o processo são necessárias para qualquer decisão sobre o negócio propriamente dito. Os autores afirmam que os problemas típicos com dados de processos de plantas industriais são:

- Absurda quantidade de dados
- Pouca confiança nos dados disponíveis
- Falta de consistência - Violam as restrições conhecidas (balanços de massa e energia)

A baixa qualidade dos dados pode resultar em tomadas de decisão baseada em dados pouco confiáveis em todos os níveis da organização e conseqüentemente prejuízos financeiros. Como todas as medições de processo estão sujeitas a algum tipo de erro, todos estes dados armazenados podem estar corrompidos de alguma maneira, seja com pequenos erros aleatórios como com grandes erros grosseiros. E assim não se pode esperar que os dados obedeçam às leis de conservação de massa ou de energia. O uso racional de grandes volumes

de dados requer a aplicação de técnicas adequadas para aumentar a sua precisão (Wang e Romagnoli, 2003). Estes erros são gerados por diferentes fontes (Wang e Romagnoli, 2003; Bagajewicz e Cabrera, 2003, Bagajewicz, Jiang e Sanchez, 2000):

- Relacionadas aos instrumentos:
    - Irreprodutibilidade
    - Degradação
    - Mau funcionamento
    - Aferição
    - Instalação imprópria
    - Falha completa
- Erros humanos
- Erros relacionados ao processo

Quando dados corrompidos são usados para estimação do estado do processo, tomada de decisões operacionais ou controle de processos, o estado verdadeiro da planta é mal representado. O desempenho resultante do controle do processo pode levar a planta a operar em um ponto de operação sub-ótimo ou inseguro, levar à perda de especificação de produtos, poluição ambiental, perdas financeiras ou altos custos de operação (Morad *et al.*, 2005). Portanto um fator muito importante na qualidade de processos, segurança e economia é a precisão das medições das variáveis de processo. Perdas materiais não detectadas podem gerar grandes efeitos nos custos operacionais, na segurança do meio ambiente e até acidentes em plantas que podem levar à perda de vidas, equipamentos e receita (Devanathan *et al.*, 2004).

As técnicas de redução de erro podem ser aplicadas a qualquer processo industrial fazendo parte de uma estratégia conhecida como *condicionamento de dados* (ou ainda retificação de dados), na qual a reconciliação de dados está incluída. Esta estratégia envolve uma série de etapas, podendo ser baseadas em modelos do processo ou não. O objetivo é tornar este enorme banco de dados do processo mais preciso, de menor dimensão, e que todas as informações relevantes estejam presentes. Isto acontece, pois se deseja poder reconstruir o estado da planta a partir das variáveis armazenadas. Na Figura 1.1 são mostradas as principais etapas e os principais tópicos relacionados ao condicionamento de dados.

Em geral, a problemática de erros em medições está associada à área da engenharia relacionada ao controle de processos, mas na realidade, a abrangência do tema deve ser levada em consideração em todos os níveis da pirâmide de gerenciamento de processos mostrada na Figura 1.2. Medições pouco precisas irão influenciar desde o nível mais baixo da pirâmide – o

controle regulatório, como os níveis superiores – planejamento e programação de produção e o gerenciamento do negócio propriamente dito. À medida que os dados de processo são utilizados nos níveis mais elevados da pirâmide, mais receita está associada à tomada de decisões baseada nestes.

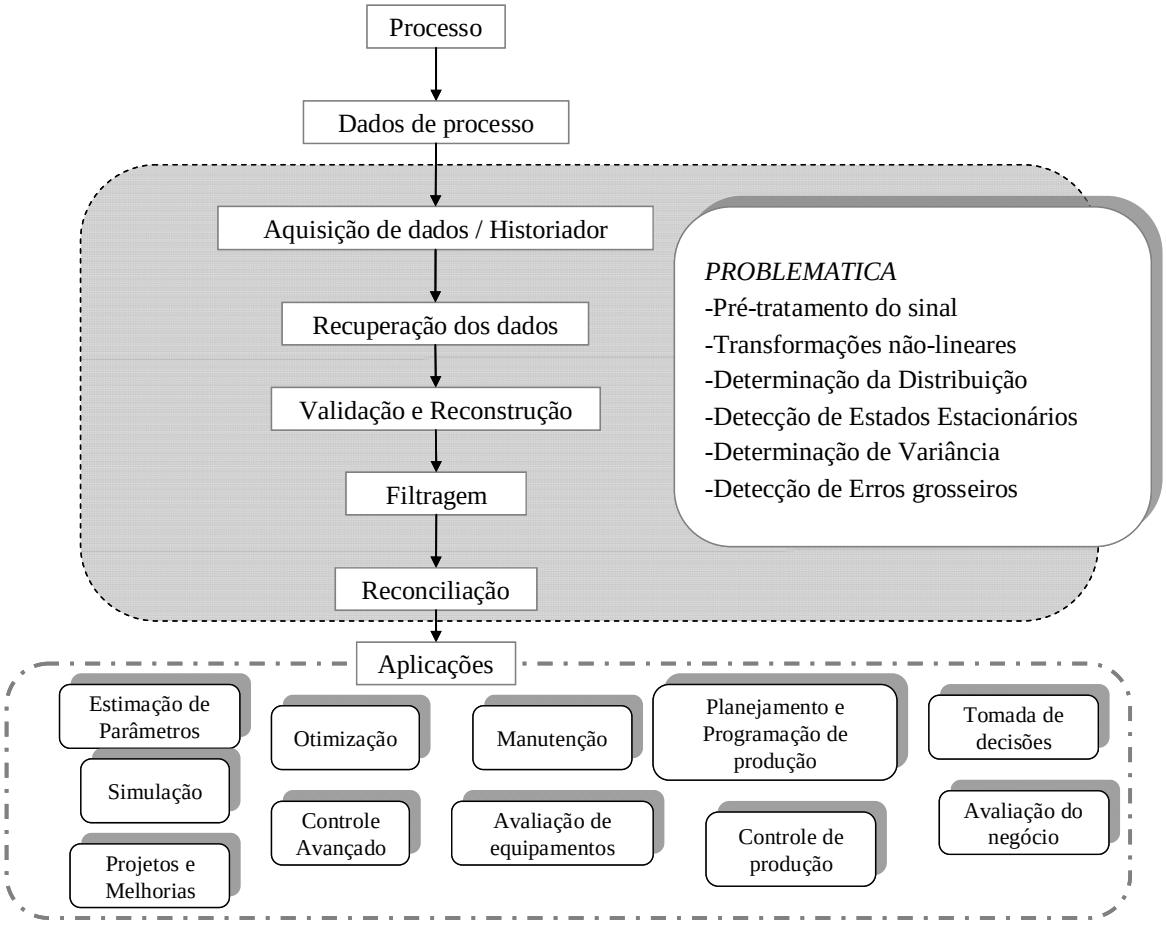

Figura 1.1: Condicionamento de dados e aplicações.

Figura 1.2: Camadas normalmente utilizadas para organizar o controle de processos (Trierweiler e Farenzena, 2007).

A motivação inicial para este trabalho vem da experiência pessoal do autor desta dissertação, em planejamento, controle e programação de produção em uma empresa petroquímica de grande porte. Como aprendizado desta experiência conclui-se que sempre é necessário o tratamento dos dados de processo, visto que é necessária a utilização de um grande volume de informações e variáveis em tarefas simples do dia-a-dia de um engenheiro químico. Sem a aplicação de uma metodologia adequada, as premissas utilizadas para o ajuste dos dados (para utilização, por exemplo, em um balanço de energia) facilmente tornam-se subjetivas e isto faz com que a tarefa de lidar com tanta informação seja um grande desafio.

Na prática, quando não se utiliza uma metodologia de reconciliação de dados, a simples tarefa de fechamento de balanço de massa é realizada utilizando-se o "*bom senso*" do engenheiro. Esta não deixa de ser uma maneira de se fazer reconciliação de dados, muito intuitiva e primitiva, mas não há garantias que os resultados obtidos sejam uma boa aproximação da realidade. Quando o engenheiro se depara com esta problemática é que aparece a necessidade de procurar ferramentas para realização da tarefa de garantir (ou melhorar) a precisão da informação proveniente do processo.

Para ilustrar este ponto de vista pode-se citar um exemplo comum do dia-a-dia de quem trabalha no topo da pirâmide mostrada anteriormente: Existe a necessidade de fechar um balanço de massa para verificação de rendimento de uma nova matéria-prima de modo a quantificar o sucesso de um negócio (por exemplo, a compra da matéria-prima). Assim o engenheiro parte do histórico das variáveis envolvidas e utiliza o que está ao seu alcance: médias temporais e desvios padrões. Quando o balanço não pode ser considerado satisfeito, nem por aproximação, então parte-se para o que se poderia chamar de "metodologia intuitiva de ajustar os dados". Uma prática vista por este autor é a utilização de heurísticas do tipo: "*Vou ajustar o valor proveniente daquele medidor que sempre marca menos um número de toneladas X e distribuir o que sobra pelos outros medidores ponderados pelas suas vazões médias*".

Esta heurística não é irreal e tem certo embasamento estatístico. Mas, além de levar em consideração um julgamento subjetivo (*aquele medidor que sempre marca menos. Mas quanto tempo o analista está acompanhando?*), não se consegue diferenciar o quanto o erro de medição está sendo minimizado ou simplesmente manipulado. Muitas vezes este procedimento nada mais é do que "inventar" valores para as variáveis e desprezar o que de mais valioso existe sobre o estado atual do processo. A idéia de aplicar uma metodologia de reconciliação de dados é ajustar os dados medidos de maneira que o erro inerente ao processo de medição seja minimizado, aproveitando a melhor informação possível sobre o atual estado do processo – *A MEDIÇÃO*.

Em geral, a redução da variação gerada por ruídos aleatórios em medidas de variáveis de processo é classificada como filtragem ou suavização (dependendo do horizonte temporal), mas quando as estimativas devem satisfazer as restrições físicas, esta tarefa passa a se chamar *reconciliação de dados* (Devanathan *et al*., 2000). A reconciliação de dados (RD) ajusta as medidas de processo fazendo com que estas satisfaçam balanços de massa ou energia sendo realizada sobre a redundância nas medidas.

A informação obtida a partir das medidas de processo em conjunto com o modelo estatístico das medições e o modelo do processo gera redundância e esta é a base para a melhoria na acuracidade das medições de processo quando a técnica de reconciliação de dados é aplicada. No sistema de reconciliação de dados, apresentado na Figura 1.3, é mostrado que o sistema permite (Benqlilou, 2004):

- ✓ O aumento da confiança na medição;
- ✓ Obtenção do valor mais provável para as variáveis não-medidas (também conhecido como problema de cooptação) incluindo parâmetros de modelo;
- ✓ Detecção de falhas em instrumentos (*bias*) e no processo (*vazamentos*);
- ✓ Determinar a localização ótima de instrumentos de medição;
- ✓ Obter uma melhor caracterização do estado atual do processo permitindo que se possa operar mais próximo da especificação;
- ✓ Estimar a eficiência dos equipamentos;
- ✓ Reduzir erros de modelagem;
- ✓ Justificar tarefas de manutenção (como por exemplo, calibração de instrumentos, limpeza de equipamentos, melhorias no processo);

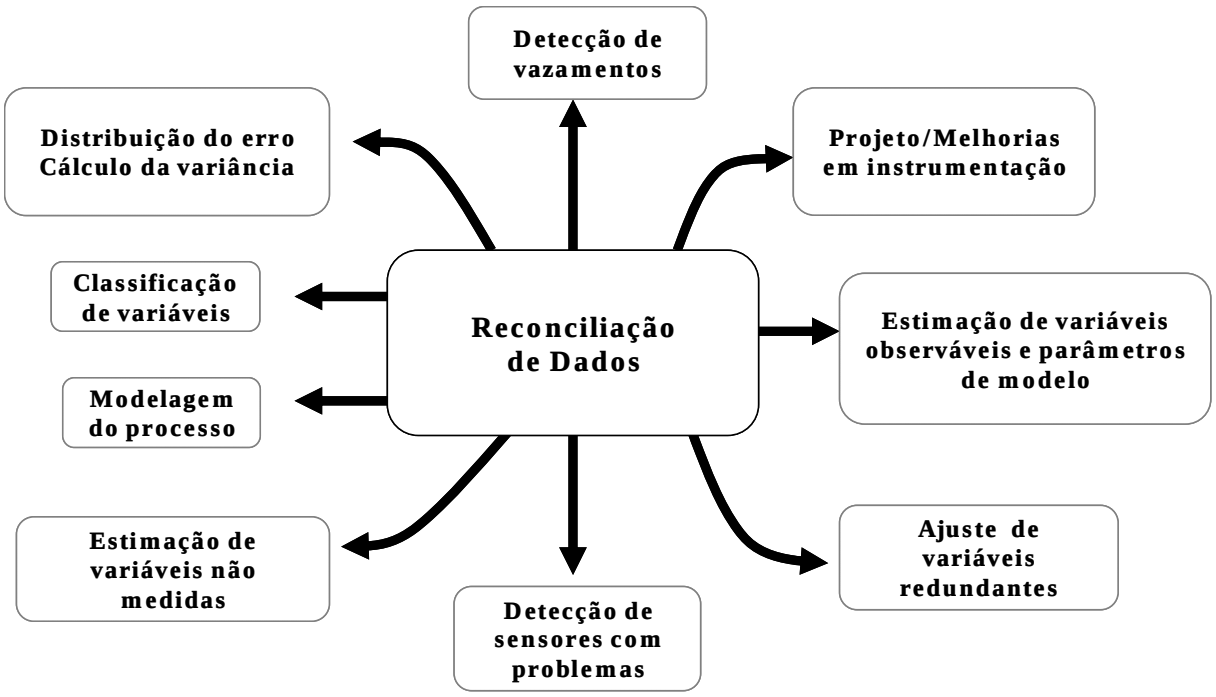

Figura 1.3: Sistema de Reconciliação de dados (adaptado de Benqlilou, 2004).

O problema de RD é geralmente tratado em três etapas:

I. Classificação das variáveis

II. Detecção de erros grosseiros

III. Reconciliação de dados propriamente dita

A etapa de classificação das variáveis serve para definir o conjunto necessário de variáveis para realizar a RD, definir se o problema pode ser reduzido de tamanho e avaliar a medição. Para isto existem diferentes métodos de classificação de variáveis na literatura e nesta dissertação é utilizado o método da decomposição QR (Sanchez e Romagnoli, 1995) devido à simplicidade e afinidade com as técnicas utilizadas.

A etapa de detecção de erros grosseiros (DEG) é crucial para o bom resultado da reconciliação de dados. Quando existem erros grosseiros, o conjunto de dados não pode ser submetido à etapa de RD (da maneira clássica) porque este serão espalhados pelo conjunto de dados. Existe uma variedade muito grande de métodos, com características diferentes e a maioria é baseada em testes estatísticos para detecção, localização e estimação dos erros grosseiros, de maneira independente à solução do problema de reconciliação. Entretanto, existem alguns métodos em que o problema de reconciliação é resolvido simultaneamente com o problema de DEG sendo tratado como um problema de estimação pura baseados em estatística robusta.

Já a etapa de reconciliação propriamente dita visa à solução de um problema de otimização e tem como objetivo estimar o estado mais provável das variáveis de maneira que satisfaçam o modelo do processo. Esta etapa esta relacionada diretamente ao tratamento do ruído do sinal de medição.

Além destas etapas o problema de reconciliação envolve outros aspectos, como por exemplo, detecção de estados estacionários, pré-tratamento dos dados, determinação da variância, avaliação da qualidade da medição (localização, aferição, tipo de instrumento) e a determinação do tipo de distribuição do ruído. Estes assuntos não serão tratados diretamente nesta dissertação, embora alguns dos tópicos sejam avaliados com base em técnicas amplamente utilizadas na literatura e muitos dos conceitos envolvidos estão implícitos nas premissas utilizadas neste trabalho e que serão apresentadas no momento apropriado.

O principal objetivo deste trabalho é realizar um estudo completo em reconciliação de dados e detecção de erros grosseiros visando à avaliação do desempenho, de maneira justa, das diferentes técnicas existentes. Dentro deste contexto são analisados diferentes parâmetros que influenciam cada etapa. Além da comparação entre as técnicas, uma nova estratégia de detecção de erros grosseiros é proposta. Esta nova estratégia é desenvolvida e validada utilizando trabalhos reportados na literatura.

1.2 Estrutura da Dissertação

Esta dissertação apresenta-se dividia em sete capítulos, conforme descritos a seguir:

O Capítulo 1 trata da introdução ao tema a ser abordado na tese. No Capítulo 2 é feita a revisão bibliográfica dos principais trabalhos publicados a respeito do tema, descrevendo de forma sucinta o que foi desenvolvido em cada trabalho e o desenvolvimento teórico dos conceitos fundamentais utilizados para a aplicação das técnicas estudadas.

O Capítulo 3 é dedicado ao desenvolvimento teórico das técnicas de reconciliação. No Capítulo 4 são apresentados os algoritmos de detecção de erros grosseiros.

No Capítulo 5 é apresentada a metodologia. Neste estão incluídos os casos utilizados, as premissas para os diferentes testes e análises, e ainda algumas considerações com relação às condições para geração dos dados.

O Capítulo 6 apresenta os resultados obtidos e é dividida em quatro seções. A primeira seção trata do problema de reconciliação e a influência do pré-tratamento de dados. Na segunda seção, são tratados os resultados obtidos para a comparação entre os diferentes métodos de detecção de erros grosseiros, com ênfase na detecção e identificação dos erros. Na terceira seção são apresentados os resultados para o conjunto reconciliação-detecção. É feita a comparação entre o desempenho dos diferentes tipos de técnicas de reconciliação aliada à detecção de erros grosseiros e são propostas modificações em alguns dos algoritmos. E na seção última é apresentado o desenvolvimento do novo método proposto bem como a sua validação.

No Capítulo 7 são enumeradas as principais conclusões deste trabalho, bem como algumas sugestões de trabalhos futuros na área.

Capítulo 2

Revisão Bibliográfica

2.1 Erros em Medidas de Processo: Modelo Estatístico da Medição

As medidas de variáveis de processos, como vazões, concentrações e temperaturas, estão sujeitas não somente a erros de medição, mas também a variabilidade do processo. Não podemos esperar que qualquer conjunto de medidas obedeça às leis de conservação de massa e energia. Os erros em medidas de processo são tratados na maioria das vezes como tendo um comportamento aleatório – isto é, ter um valor esperado de zero e uma variância conhecida – mas ocasionalmente são encontrados erros constantes e de maior magnitude com valores esperados diferentes de zero. O erro total na medição é definido como a diferença entre o valor real e o valor medido, e esta diferença pode ser convenientemente representada como o somatório das contribuições de dois tipos de erros: erros aleatórios (ε) e erros grosseiros (δ).

$$e = x - y = \varepsilon + \delta \tag{2.1}$$

2.2 Erros Aleatórios

O termo erro aleatório implica em que nem a magnitude, nem o sinal do erro podem ser preditos. Em outras palavras, se a medida é repetida com o mesmo instrumento de medição, sob condições externas idênticas, um valor diferente pode ser obtido dependendo do resultado do erro aleatório. Estes erros podem ser causados por um número grande de fontes, como por exemplo: flutuações no fornecimento de energia, ruído na rede de transmissão, ruído na conversão do sinal, filtros, mudanças nas condições ambientais, ruído térmico nos componentes eletrônicos, etc.

Erros aleatórios não podem ser completamente eliminados e estarão sempre presentes em qualquer medição. O erro aleatório está relacionado às componentes de alta freqüência do

sinal medido, sendo pequeno em magnitude, exceto por picos ocasionais (Romagnoli e Sanchez, 2000). O efeito dos erros aleatórios nas medições é modelado como um somatório de contribuições. Se considerarmos que os erros nas medições são compostos por distúrbios aleatórios ("ruídos"), provenientes de diferentes fontes e que o número destas fontes é suficientemente grande, *o teorema do limite central* se aplica (Bagajewicz, 1996), isto é, a soma de todos estes distúrbios tende à distribuição normal.

Uma observação importante a ser feita é em relação à nomenclatura do chamado "*ruído branco",* muitas vezes associada aos erros aleatórios. Este aparece em vários trabalhos de aplicação das técnicas de reconciliação (sem desenvolvimento teórico) e esta definição não é propriamente adequada. O ruído branco não necessariamente é Gaussiano nem uniforme. O ruído branco é um sinal aleatório com densidade de potência espectral uniforme, isto é, um sinal que contém potência igual dentro de uma banda de freqüência centrada em qualquer freqüência. Nesta dissertação será usado o caso particular mais comum que é o *ruído branco Gaussiano.*

Em Madron et al. (1992), as flutuações são classificadas em duas categorias: aleatórias e determinísticas. As flutuações determinísticas seriam as geradas a partir de malhas de controle mal projetadas, por exemplo. Nesta dissertação serão consideradas somente as flutuações aleatórias e receberão o tratamento dado às variáveis aleatórias gaussianas contínuas.

2.2.1 Modelo estatístico da medição na presença de erros aleatórios

Definindo-se o erro aleatório como uma variável aleatória:

$$\varepsilon_i = y_i - x_i \tag{2.2}$$

onde y é o valor medido, x é o valor verdadeiro e ε é o erro aleatório. E assim pode-se desenvolver o *modelo estatístico da medição* na presença de erros aleatórios como sendo:

$$y = x + \varepsilon, \quad \varepsilon \sim N(0, V) \tag{2.3}$$

onde $\varepsilon \in \Re^n$ é o vetor dos erros aleatórios caracterizado pela matriz de variância V. Se na medição estão presentes somente erros aleatórios, para cada componente x_i de x, a seguinte função distribuição de probabilidade (pdf) pode ser definida:

$$p_i(y_i|x_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\left(\frac{y_i - x_i}{\sigma_i} \right)^2 \right) \tag{2.4}$$

onde σ_i^2 são os elementos da diagonal de V e p é a pdf normal. Desta forma, surge a premissa mais importante - *Os erros aleatórios seguem uma distribuição Gaussiana, com média zero e variância conhecida V e, conseqüentemente, as medidas de processo também seguem esta distribuição*. A Função densidade de probabilidade é ilustrada na Figura 2.1.

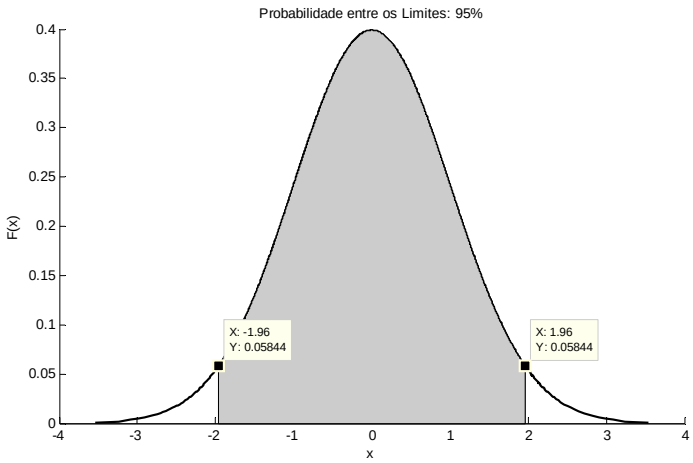

Figura 2.1: Função densidade de probabilidade normal padrão.

Como conseqüência disto, a variância dos erros aleatórios segue uma distribuição qui-quadrado. Esta é uma transformação conhecida da distribuição normal. Considerando ε_1, ε_2,..., ε_υ, variáveis independentes aleatórias descritas pela distribuição normal padrão, então a variável qui-quadrada é definida como:

$$\chi^2(\nu) = \sum_{i=1}^{\nu} \varepsilon_i^2 \tag{2.5}$$

onde υ são os *graus de liberdade*. Assim, a sua função distribuição é representada analiticamente na Equação 2.6 (onde Γ é a função gama). A distribuição qui-quadrado é ilustrada na Figura 2.2:

$$F\left(\chi^2\right) = \frac{\left(\chi^2\right)^{(\nu-2)/2} \exp\left(\frac{\chi^2}{2}\right)}{2^{(\nu/2)}\left[\Gamma(\nu/2)\right]} \tag{2.6}$$

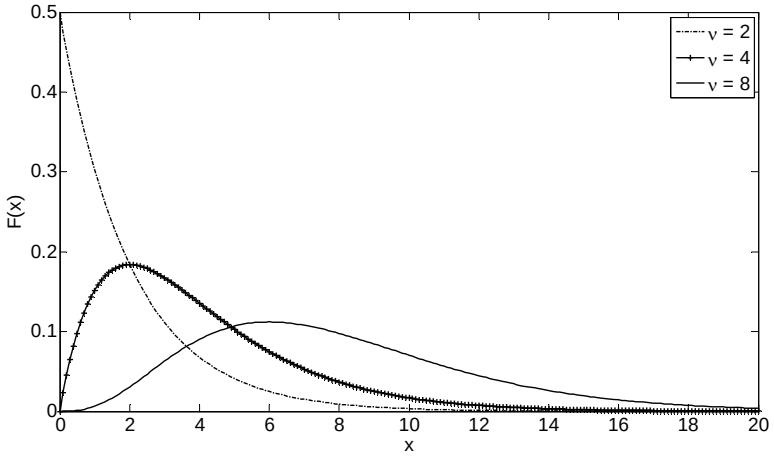

Figura 2.2: Função densidade de probabilidade qui-quadrado.

2.2.2 A Distribuição Gaussiana e suas implicações

O objetivo desta seção é mostrar um pouco da discussão existente entre diversos autores das áreas de RD e DEG, sobre a escolha da distribuição dos erros aleatórios. A premissa de que os erros aleatórios seguem uma distribuição normal é a mais utilizada na literatura e, por ser muito forte, é constantemente questionada.

Existe uma série de trabalhos que tratam sobre o assunto, pois muitas vezes os autores se depararam com situações em que esta premissa não é válida. Por exemplo, nos casos de medidores de vazão do tipo placa de orifício (como em Bagajewicz, 1996), ou ainda, na tentativa de estender as técnicas já consolidadas em reconciliação linear para problemas não lineares. Assim, é encontrada na literatura uma série de trabalhos que tem por objetivo avaliar a possibilidade de utilizá-la em um momento inicial sem que toda a metodologia de tratamento de dados tenha que ser revalidada/reaplicada caso os dados indiquem o contrário.

A discussão é muito interessante, visto que, para os mais céticos, tal distribuição não existiria na prática (Madron et al., 1992; Mah, 1990; Wang e Romagnoli, 2003; Bagajewicz, 1996), mas as conseqüências matemáticas, algébricas e simplificadoras da utilização de tal premissa são interessantes. Em Mah (1990), o autor reconhece que esta premissa é utilizada devido à necessidade de informação sobre a distribuição dos dados e que aparenta ser interessante em um primeiro momento, a menos que os dados indiquem fortemente o contrário. Madron et al. (1992) reforça esta posição e adiciona que no caso da distribuição não ser conhecida, ao assumir-se a distribuição gaussiana, esta premissa adiciona mínima informação ao problema de RD e não sendo necessárias correções posteriores.

Em Wang e Romagnoli (2003), os autores afirmam que a distribuição normal não existe na prática real de engenharia química e que esta premissa é difícil de ser garantida mesmo para medições de alta qualidade. Bagajewicz (1996) argumenta que o processamento de sinais eletrônicos normalmente distribuídos envolve, muitas vezes, transformações não-lineares na rede de instrumentos gerando distribuições não normais.

Madron (1992) reconhece a dificuldade de se garantir a consistência, afirma que não existem medidas com distribuição gaussiana, mas ignora a própria afirmação com base na seguinte justificativa:: *"... erros são usualmente pequenos e muitas funções podem ser vistas como aproximadamente lineares dentro de uma região limitada."*. Após, o autor apresenta três justificativas para a escolha da distribuição normal, da seguinte maneira:

I. É sabido que a distribuição normal aproxima bem o comportamento de medidas nas ciências naturais, particularmente dentro do desvio médio de $\pm 3\ \sigma$.

II. Um erro é geralmente um somatório de um grande número de erros elementares. De acordo com o teorema do limite central, sob certas condições aceitáveis, a distribuição deste tipo de somatório aproxima a distribuição normal (para um grande número de erros elementares).

III. O modelo teórico da função distribuição normal é bem desenvolvido e é fácil de tratar matematicamente. Os valores para a função de probabilidade para a distribuição normal padrão estão disponíveis em formas de tabela em qualquer livro de estatística facilitando a solução de problemas práticos.

Engenheiros químicos têm sido treinados a pensar sobre erros em medições sob ponto de vista de teoria de controle. Dentro desta, a magnitude dos erros é importante e poucas vezes a sua distribuição é considerada. A distribuição dos erros não é distorcida quando as transformações são lineares. Entretanto, o processamento do sinal é não-linear e como conseqüência as distribuições de probabilidades são distorcidas (Bagajewicz, 1996).

Neste sentido, existem alguns trabalhos que avaliam as conseqüências da utilização desta distribuição e adequações necessárias. Um dos principais trabalhos é o de Crowe (1996), onde foi realizado um estudo utilizando teoria da informação (maximizando a entropia da informação) para obter as distribuições de probabilidade que incorporariam mínima restrição devido à sua utilização. Neste, o autor conclui que a premissa da distribuição normal multivariada é a que incorpora a mínima informação ao sistema desde que o erro entre a medida e seu valor verdadeiro seja de até 30% e a variância das medidas seja conhecida. Para conjunto de dados com a variância desconhecida o autor afirma que a melhor premissa é assumir uma distribuição normal multivariada truncada (uma distribuição gaussiana com suporte finito).

Já em 1996, Bagajewicz realizou um estudo sobre a relação entre a premissa de distribuição normal com transformações não-lineares no sinal de medição. O autor usou como exemplo o caso, já citado, de medidas de vazão provenientes de medidores do tipo placas de orifício. Assim, neste trabalho, é sugerido que para variáveis diretamente medidas (temperatura, composição,...) a premissa é adequada, mas quando isto não ocorre, e são necessárias transformações, a distribuição de probabilidade é distorcida e perde sua validade. Por isto foi proposta uma formulação alternativa para o problema de reconciliação, onde são incluídas no problema de RD as equações de transformação da medida primária em medidas secundárias, e as transformações não-lineares são linearizadas visando esta inclusão.

Neste contexto, tanto a reconciliação de dados como a detecção de erros grosseiros tem evoluído de maneira que os pesquisadores também procuram técnicas onde não exista a necessidade de se garantir determinada distribuição para os erros de medição. Algumas destas técnicas serão apresentadas mais adiante neste trabalho.

2.3 Reconciliação de dados

Nos últimos 50 anos, tanto a reconciliação de dados como as técnicas de detecção de erros grosseiros têm recebido muita atenção por parte do mundo acadêmico. A principal motivação dos pesquisadores é a obtenção de métodos para aumentar a precisão de dados de processo medidos. Com a melhoria e o avanço das técnicas computacionais, assim como a aquisição de dados industriais, é cada vez mais possível tratar uma quantidade *enorme* de dados de processo. Isto fez com que a maioria dos pesquisadores buscasse nas técnicas

estatísticas e probabilísticas a base para os métodos desenvolvidos na área. O problema de reconciliação de dados é tratado como um problema de minimização dos ajustes feitos nas medições. Este problema tem como restrições modelos do processo e estes podem ser:

- ✓ Lineares (balanço material)
- ✓ Bilineares (balanços por componentes sem reação química)
- ✓ Não-lineares (balanço energético, reações químicas,...)

E ainda podem ser

- ✓ Estacionários
- ✓ Dinâmicos

Na Tabela 2.1 é mostrado um panorama geral das técnicas mais utilizadas, não só para reconciliação linear, mas mostrando as técnicas utilizadas para as soluções dinâmicas e também as não-lineares visando situar o leitor em um contexto mais amplo. Todas as técnicas citadas, e que não fazem parte do universo deste trabalho, poderiam ser utilizadas, mas nem todos os métodos aqui apresentados foram estendidos para a aplicação em problemas mais complexos, como por exemplo, reconciliação não-linear. Isto se deve muito ao fato de não existir relações óbvias entre uma função distribuição de probabilidades e transformações não-lineares destas.

Tabela 2.1: Técnicas utilizadas para Reconciliação de Dados

Problema	***Estacionário***			***Dinâmico***	
Restrição	***Linear***	***Bilinear***	***Não-linear***	***Linear***	***Não-linear***
Técnicas	Solução Analítica	Matriz de Projeção	*QP*	Filtros de Kalman	Filtros de Kalman estendidos
	QP	Solução aproximada	*SQP*	*MHE*	Solução Bayesiana
	SQP	Decomposição *QR*	Linearização	Redes Neuronais	Redes Neuronais
	Reconciliação Robusta	Linearização	*MINLP*	Solução Bayesiana	*NDDR*
	Matriz de Projeção	*QP*	Redes Neuronais	Reconciliação Robusta	Filtros Particulados
	Decomposição QR	*SQP*	Reconciliação Robusta	*Wavelets*	
	MILP	*MILP*	Algoritmos Genéticos		
	Error in Variables	Reconciliação Robusta		*DDR*	

As técnicas utilizadas para a solução do problema de RD variam de acordo com o tipo de restrição e o tipo de relação temporal. As técnicas de otimização utilizadas podem ser simples, como solução analítica do problema de mínimos quadrados sem restrições, até técnicas mais complexas, como Filtros de Kalman Estendidos, Estimador de Horizonte Móvel (MHE), Filtros Particulados e *Wavelets*.

Como já foi dito no capítulo introdutório deste trabalho, dentro do sistema de reconciliação, é necessário tratar previamente o problema de detecção de erros grosseiros. As técnicas de RD podem ser separadas ainda em 2 grandes grupos:

- *Reconciliação de dados clássica:* Onde o problema de DEG é tratado separadamente com auxílio de algoritmos para a prospecção/estimação de erros grosseiros.

- *Reconciliação de dados simultânea com a DEG*: Onde o problema de detecção é tratado simultaneamente, seja pela solução do problema de otimização da RD sendo resolvido simultaneamente com o problema de estimação de EG's (por exemplo, "*Error in Variables*"), ou ainda pela utilização de função objetivo que rejeita a presença de EGs (*Reconciliação Robusta)*

Na reconciliação de dados clássica, a forma típica da função objetivo é a quadrática dos ajustes das medidas, sendo um somatório de quadrados ponderados. Este método é conhecido como *Mínimos Quadrados Ponderados* e foi proposto em 1795 por Gauss (1809) e publicado pela primeira vez por Legendre, em 1805. A primeira aplicação do método foi para a determinação de órbita de cometas (Legendre, 1805).

Apesar do método mais utilizado para a reconciliação de dados ser do século 18, os primeiros trabalhos em reconciliação de dados datam da década de 60 do século 20. Um breve histórico do desenvolvimento das técnicas de reconciliação de dados aplicadas em engenharia química é dado a seguir, em ordem cronológica, e esta tem por objetivo esclarecer como se deu o desenvolvimento até o momento atual. Assim, a primeira solução proposta para o problema de reconciliação linear de dados foi apresentada por Kuehn e Davidson (1961). O trabalho apresentava a solução para um sistema com todas as variáveis medidas. Já em 1963, tendo como base o trabalho de Kuehn e Davidson, foi apresentado o primeiro caso de aplicação industrial (Reilly *et al.*, 1963).

Em seguida, Vaclavek (1969, 1972, 1973, 1976) apresentou uma série de publicações sobre os princípios básicos da reconciliação de dados explorando a topologia do processo para reduzir o problema geral em um problema menor. A ideia é eliminar variáveis não medidas pela combinação de unidades ou eliminar unidades que não tenham a corrente de alimentação medida. Este método é conhecido como *Esquema de Balanços Reduzidos ("Reduced Balance Scheme")*. O autor também trata das definições para os conceitos de observabilidade e redundância das variáveis.

Em 1975, Mah et al. apresentou a relação entre a álgebra e a teoria gráfica e, com o objetivo de reduzir o tamanho do problema de reconciliação, propôs um método para separar o problema geral de reconciliação em dois subproblemas.

Em 1983, Crowe et al. apresentaram um método para separar as variáveis medidas das não medidas utilizando uma matriz de projeção baseada no algoritmo de *fatoração QR – O método da matriz de projeção*. Assim, foi tratado o problema da classificação das variáveis, que é um passo importante para redução do tamanho do problema geral de estimação, incluindo-o no algoritmo de decomposição das variáveis medidas e não medidas do processo. Já em 1986, Crowe et al. expandiram o método da matriz de projeção para lidar com restrições bilineares.

Em 1988, Yamamura et al. (1988) propuseram uma função objetivo baseada no critério de informação de Akaike para identificar instrumentos com falhas.

Em 1991, Tjoa e Biegler propuseram uma metodologia de reconciliação de dados simultânea com a detecção de erros grosseiros, onde é apresentado um modelo diferenciado para a medição. Nesta, sugere-se que o modelo para EG seja também uma variável aleatória com distribuição normal e média diferente de zero. Assim, o modelo de medição passa a ser a mistura de duas distribuições gaussianas e gera-se uma função objetivo que leva em conta a presença e distribuição do erro grosseiro. A função objetivo passa a conter dois termos: um representando os erros aleatórios e o outro representando erros grosseiros, multiplicados pela respectiva probabilidade de sua ocorrência. A distribuição resultante foi chamada, em um primeiro momento de *bivariate normal distribution*, mas após a reivindicação de Mah (1997), os autores reconheceram (Biegler, 1997) que esta nomenclatura estava errada e passaram a chamá-la de *contaminated distribution.* A maior vantagem deste procedimento é a eliminação do procedimento combinatório em busca de erros grosseiros.

Em 1993, Terry e Himmelblau e em 1994, Karjala e Himmelblau, propuseram a utilização de redes neuronais para a solução do problema de RD. A utilização de redes neuronais geralmente é realizada em duas etapas. Uma etapa de projeto e ajuste da rede, tendo como desafio a aquisição de um padrão representativo de dados reconciliados para o ajuste (Yang-Guang, Thibault e Hodoui, 1997) e uma segunda etapa de solução do problema propriamente dito. A principal aplicação seriam problemas não-lineares e problemas dinâmicos, mas o interessante é a não necessidade de se supor uma distribuição dos erros *a priori*.

Em 1995, Johnston e Kramer (1995) estabeleceram uma analogia entre estimação de máxima verossimilhança e regressão robusta baseados em estatística robusta e a habilidade de rejeitar *outliers* (Huber, 1981) e os autores reportaram uma performance superior dos estimadores robustos em relação ao problema de RD convencional quando os dados continham erros grosseiros. Esta abordagem do problema de RD é chamada de *reconciliação robusta* e nesta formulação a função objetivo da formulação de mínimos quadrados é substituída por outras funções dos resíduos das restrições. A principal etapa da estimação

robusta é a escolha da função de influência (a derivada da função objetivo em relação às medições das variáveis de processo), que definirá qual a família de funções a ser usada.

Em 1997, Safavi et al. propuseram a utilização de ondaletas (*wavelets*) como técnica de estimação da densidade dos dados e solução do problema de reconciliação.

Em 2000, Soderstrom et al. apresentaram uma aplicação industrial de grande escala de uma estratégia de reconciliação dinâmica de dados baseada no trabalho de *Liebman et al.* (1992), usando a técnica de programação não-linear para resolver o problema de reconciliação e estimação. Este trabalho foi importante para demonstrar que é possível a utilização de reconciliação dinâmica em problemas de grande porte, pois esta técnica ainda é vista como computacionalmente muito complexa para ser utilizada na indústria (Bagajewicz, 2003) e é uma das razões que levam muitos pesquisadores não darem por concluídos os desenvolvimentos em reconciliação estacionária.

Romagnoli e Sánchez (2000) propuseram a utilização de uma série de famílias de função objetivo para solução simultânea do problema de RD e DEG usando reconciliação robusta.

Como os processos realmente nunca estão em estado estacionário, uma das grandes críticas para a reconciliação estacionária é de que não existem somente erros aleatórios nos dados, mas também são incluídas variações relacionadas com o processo. E quando são realizadas as médias dos dados, o espalhamento destes erros acontece de maneira imprevisível (Bagajewicz, 2003). Por isto, Bagajewicz et al. (2000) compararam o desempenho entre a abordagem integral para a reconciliação de dados dinâmica e a reconciliação de dados estacionária. Foi mostrado que na ausência de *bias* e vazamentos, o desempenho em ambas as abordagens é similar e, uma vez que a variância apropriada seja escolhida, ambos os métodos são idênticos na ausência de termos de acúmulo. Finalmente, foi feita uma análise das discrepâncias como função do acúmulo onde foi demonstrado que pode existir uma diferença muito pequena entre as soluções.

Arora and Biegler (2001) propuseram a utilização de uma forma modificada da função *fair*, não convexa, chamada de *redescending estimator*. No procedimento é utilizada uma reconciliação prévia utilizando a função *fair* tradicional, proposta por Albuquerque e Biegler (1996) como chute inicial e é realizada a sintonia do estimador utilizando o critério de informação de Akaike.

Em 2001, Soderstrom, Himmelblau e Edgar propuseram a utilização de uma técnica MILP para solucionar o problema de RD simultaneamente com o problema de DEG. A função objetivo é modificada e são adicionadas restrições que incorporam a DEG na estratégia de otimização global, de modo que cada medição seja associada a uma variável binária que define a presença do EG.

Em 2003, Wang e Romagnoli propuseram uma metodologia baseada na não idealidade da distribuição dos dados para o tratamento do problema de reconciliação de dados na

presença de erros grosseiros. Isto é feito pela combinação de programação não-linear e princípios de máxima verossimilhança após a distribuição do erro de medição ser devidamente caracterizada (Wang e Romagnoli, 2003). Nesta abordagem o problema de reconciliação é posto como um problema de otimização e se leva em conta uma abordagem probabilística para o modelo do erro de medição. Assim, o erro total da medição é modelado como sendo composto de duas partes: uma distribuição estreita representando o ruído de medição e uma mais larga representando o erro grosseiro.

Ainda existem trabalhos que propõem que o tratamento de erros aleatórios seja feito utilizando reconciliação via a técnica de estimação Bayesiana (Johnston e Kramer, 1995, 1998; Chen et al., 1996). Nesta, primeiro a distribuição dos erros é definida e depois a forma do problema de reconciliação é definida. Este método utiliza programação não-linear e princípios de máxima verossimilhança após ter a distribuição dos erros devidamente caracterizada. O erro, em geral, é composto por duas partes: uma distribuição mais estreita representando o ruído na medição e outra mais larga caracterizando erros grosseiros. As duas distribuições são então pesadas pela probabilidade de existir ou não erros grosseiros nas medidas. Uma vantagem é que este método já condensa a solução do problema de reconciliação com o algoritmo de detecção de erros grosseiros. As desvantagens são que a escolha/determinação das distribuições ainda é problemática e a probabilidade com a qual as distribuições serão pesadas deve ser definida previamente. Esta difere da metodologia apresentada por Tjoa e Biegler (1991), pois a distribuição do erro grosseiro não é completamente imposta pelo modelo e pode ser modificada caso os dados não confirmem a distribuição suposta.

Neste trabalho são apresentadas as técnicas de reconciliação clássica, utilizando a solução analítica aliada à decomposição QR para o tratamento de variáveis não medidas e a reconciliação robusta, onde se utilizada programação quadrática seqüencial (SQP) e diferentes funções objetivo. Além disto, é avaliado o impacto do pré-tratamento de dados e de diferentes métodos de detecção de erros grosseiros na qualidade final dos dados estimados.

2.4 Detecção de Erros Grosseiros

2.4.1 Erros Grosseiros

Os erros classificados como sendo grosseiros podem ser tratados tanto como variáveis não aleatórias, como variáveis aleatórias com uma determinada distribuição. A definição mais simples sobre o comportamento deste tipo de erros é a seguinte: *"Se a medição é repetida com o mesmo instrumento, sob as mesmas condições, a contribuição sistemática do erro grosseiro no valor medido será a mesma"* (Narashiman e Jordache, 2000).

Os erros grosseiros podem ser causados por eventos não aleatórios como:

- ✓ Mau funcionamento de instrumentos
- ✓ Mudanças nas condições operacionais

- ✓ Instrumentos operando fora das condições em que foram calibrados ou fora de sua faixa de medição
- ✓ Distorção no perfil de escoamento causada por bolhas de gás em correntes líquida
- ✓ Instalações impróprias do instrumento de medição
- ✓ Falta de calibração
- ✓ Falha completa do instrumento
- ✓ Depósitos de sólidos
- ✓ Vazamentos
- ✓ Dados de laboratório (erro humano)

Ou ainda podem ser causados por eventos aleatórios como, por exemplo:

- ✓ Deterioração de partes mecânicas
- ✓ Falhas eletrônicas
- ✓ Descargas elétricas
- ✓ Corrosão ou desgaste de sensores
- ✓ Falhas completas de instrumentos

Erro grosseiro ocorre menos freqüentemente que o erro aleatório, mas a sua magnitude é maior. Segundo Romagnoli (2000), seguindo bons procedimentos de instalação e manutenção de sensores, é possível garantir que erros grosseiros não estejam presentes na medição pelo menos em algum momento. Erros grosseiros causados por descalibração de instrumentos podem ocorrer em certo instante particular e após assumir uma magnitude constante. Outros erros grosseiros causados por desgaste ou deposição de sólidos podem ocorrer gradualmente em um período de tempo e a magnitude deste erro aumentar ao longo do tempo vagarosamente por um período longo. Este tipo de erro grosseiro pode passar despercebido por anos (Upp e LaNasa, 2002).

2.4.2 Caracterização do erro grosseiro

Os erros grosseiros são separados em 2 categorias devido à diferença no tratamento matemático: podem ser do tipo *bias* ou do tipo vazamento. Erros grosseiros do tipo bias são modelados como um termo a ser somado no valor medido da variável:

$$y_i = x_i + \varepsilon_i + \delta_i \tag{2.7}$$

onde y_i é o valor medido, x_i é o valor verdadeiro da variável, ε_i é o erro aleatório e δ_i é o erro grosseiro. Na literatura é mais comum que os erros grosseiros do tipo bias sejam tratados como um desvio da média como ilustrado na Figura 2.3.

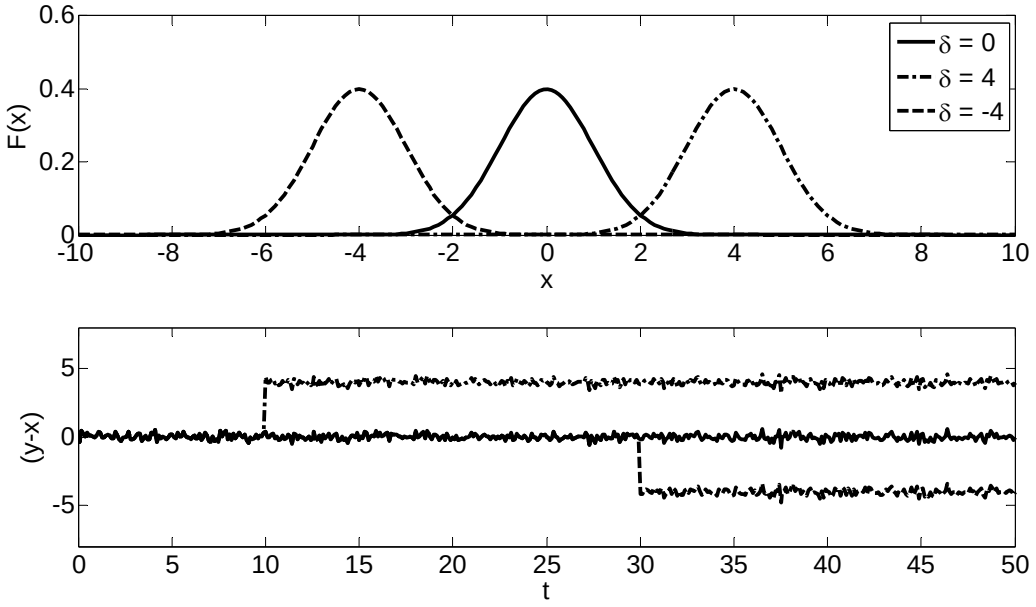

Figura 2.3: Exemplo de "bias"

Já os erros grosseiros do tipo vazamento são modelados na restrição do processo. Supondo um balanço material, onde A é a matriz de incidência das restrições (sinal positivo quando é uma corrente de entrada e negativo, caso contrário), as restrições serão do tipo:

$$Ax = 0 \tag{2.8}$$

Na presença de vazamentos as restrições são modeladas:

$$Ax + bm = 0 \tag{2.9}$$

onde *b* é o vetor unitário índice do nó com suspeita de vazamento e *m* é a sua magnitude.

Quando um erro grosseiro está presente em um conjunto de dados a ser reconciliado, o procedimento de reconciliação não apresentará bons resultados visto que estes erros serão espalhados por todo o conjunto de dados. Este espalhamento é também chamado de *smearing* (Liebman et al, 1992). Por isto, é muito importante que estes erros sejam identificados, compensados ou eliminados previamente. A Figura 2.4 mostra o efeito de smearing após os dados contendo erros grosseiros serem reconciliados, para uma rede de trocadores com reciclo, ilustrada na Figura 2.5.

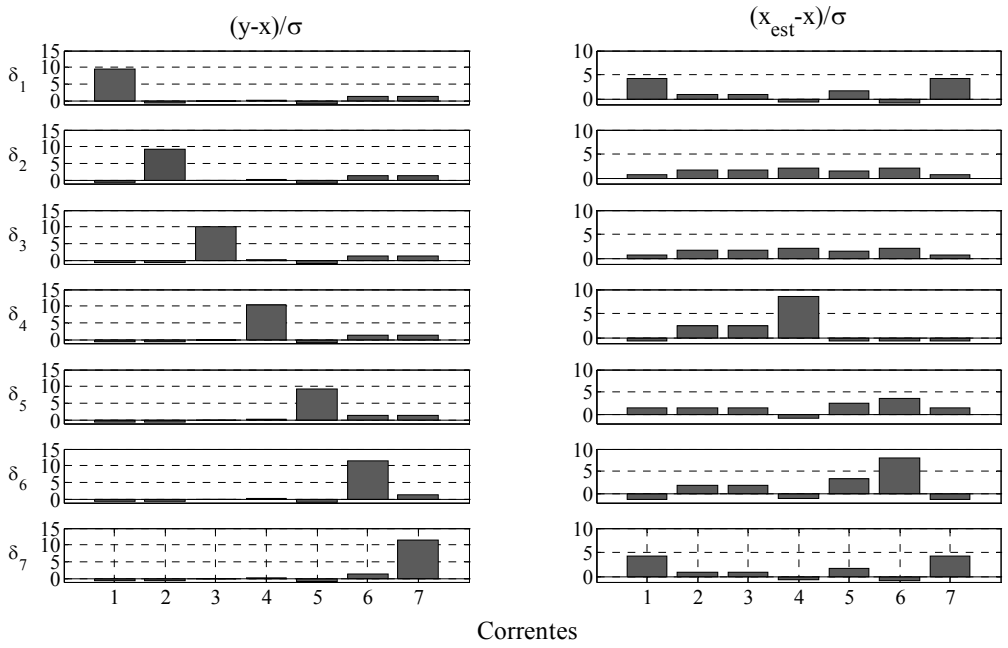

Figura 2.4: Efeito "smearing" para erro grosseiro adicionado em cada corrente de processo.

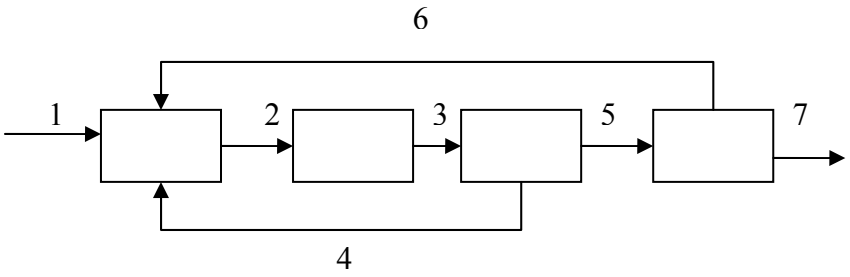

Figura 2.5: Rede de trocadores com reciclo (Rosenberg et al., 1986).

Na Figura 2.4, a coluna da esquerda mostra o erro entre o valor medido e o valor verdadeiro da variável e na segunda coluna é o erro entre o valor reconciliado e o valor verdadeiro da variável. Assim pode-se verificar que o espalhamento do erro grosseiro é diferente para cada variável, dependendo fortemente da topologia do processo. Por causa deste efeito, Bagajewicz (2007) define acuracidade como um somatório entre a precisão da medida e o bias induzido no processo de reconciliação. O autor apresenta diferentes tipos de modelagem para erros grosseiros apresentados na Figura 2.6. O autor propõe uma série de modelos matemáticos relacionados com os tipos de falhas dos instrumentos, mas com o enfoque de avaliar a evolução destas falhas ao longo do tempo e definir uma metodologia para calcular a acuracidade dos instrumentos. O autor separa os erros grosseiros do tipo bias em 3 categorias:

I. Erros grosseiros com valor constante: geralmente causados por falhas de componentes eletrônicos, devido às condições ambientais não favoráveis (alta

umidade, alta temperatura) ou à presença de efeitos elétricos (raios, mudança na impedância de resistores, problemas de aterramento)

II. Bias aleatórios do tipo rampa ou não-lineares: causados por desgaste ou corrosão, deterioração de partes mecânicas e algumas falhas eletrônicas. Podem ser rampas de desgaste ou apresentar perfil côncavo (corrosão que diminui ao longo do tempo devido à camada protetora) ou ainda convexo (degradação mecânica).

III. Bias do tipo determinístico: não são eventos aleatórios, mas processos contínuos como, por exemplo, depósito de partículas em sensores em contato com o elemento primário de medição, erosão em tubulações gerando variação na rugosidade, bloqueios de tomadas de pressão, mudanças nas dimensões originais da tubulação, mudanças no perfil de escoamento. Em geral podem chegar a um equilíbrio após certo tempo – como no caso em que partículas se depositam na superfície – geram um perfil assintótico.

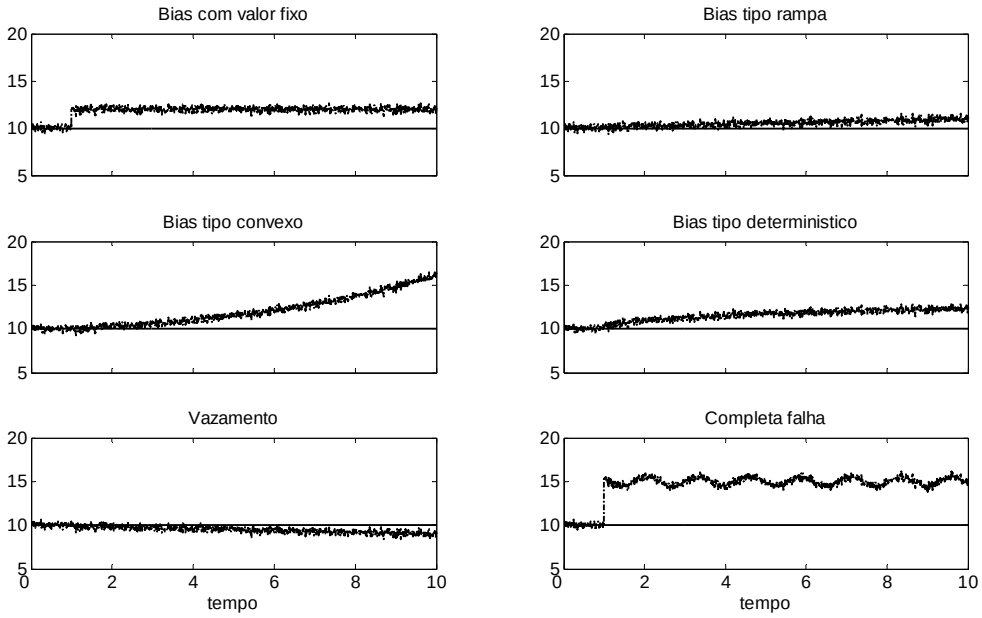

Figura 2.6: Exemplos de diferentes tipos de erros grosseiros.

Como a finalidade deste trabalho não é a detecção do tipo de falha dos instrumentos, e sim a retificação dos dados, o modelo utilizado para EG do tipo bias será a da mudança de média, visto que na RD estacionária, os dados são provenientes de médias e esta aproximação é suficiente para a detecção de todos os outros tipos de erros.

2.4.3 Métodos de detecção de Erros Grosseiros

Se todas as medidas forem ajustadas para atender as leis de conservação, na presença de erros grosseiros, as estimativas resultantes serão afetadas pelo espalhamento deste erro por todo o conjunto de dados. Dados reconciliados, na presença de erros grosseiros não serão um indicativo confiável do estado do processo. O tratamento matemático de variáveis com erros grosseiros chama-se *Detecção de Erros Grosseiros (DEG).*

Erros grosseiros devem ser identificados, corrigidos ou as medidas devem ser descartadas. Uma estratégia para a detecção de erros grosseiros deve ter a habilidade de:

- Detectar a presença de um ou mais erros grosseiros
- Identificar o tipo e a localização do erro grosseiro
- Identificar múltiplos erros grosseiros
- Prover estimativas para os erros grosseiros

As técnicas de detecção de erros grosseiros clássicas baseiam-se na aplicação de testes estatísticos no conjunto de dados, para determinar se as medições seguem uma distribuição com média zero. Assim aplica-se um teste de hipóteses na média dos dados do tipo:

H_0: Hipótese nula: $\mu = 0$

H_1: Hipótese alternativa: $\mu \neq 0$

Onde o teste é:

$$t = \frac{\hat{E}(y_i - x_i)}{\hat{\sigma}} \tag{2.10}$$

onde o sinal ^ representa a estimativa para o valor esperado do erro de medição (y_i - x_i) e a estimativa da variância dos dados V (Ozyurt e Pike, 2004). Este teste é a base para todos os procedimentos de DEG clássicos. Se uma função densidade de probabilidade pode ser considerada para *t*, grandes valores de *t* descreverão situações menos esperadas e darão a prova para que H_1 seja verdadeira, i.e., a existência de erros grosseiros.

Diversos testes estatísticos foram construídos para a detecção de erros grosseiros, de maneira que fossem exploradas as relações entre o conjunto de medições e o modelo do processo. Os testes mais comuns utilizam as relações entre os dados medidos e o resíduo das restrições de balanço ou ainda entre os ajustes realizados em uma reconciliação prévia dos dados. A seguir é apresentado um histórico da evolução das técnicas para detecção de erros grosseiros, assim como uma breve descrição.

O primeiro caso de aplicação industrial da DEG foi apresentado em 1963 por Reilly e Carpani. Neste foram apresentados dois dos mais populares testes de detecção de erros grosseiros - *O Teste Global (GT)* e o *Teste Nodal (NT)*. Ambos os métodos são baseados nos resíduos do modelo de restrição. O objetivo do GT é determinar se todo o conjunto de dados segue a distribuição normal comparando a variância dos resíduos das restrições com a variância esperada caso o conjunto de dados seguisse uma distribuição normal padrão. Já a premissa do NT é avaliar o quão bem as medições satisfazem cada uma das restrições utilizando os resíduos das restrições para detectar a existência de erros grosseiros. A maior

desvantagem para ambos os testes GT e NT é que, apesar deles poderem determinar se existem erros grosseiros no conjunto de dados, estes são incapazes de indicar qual a medição que contém erro. Por causa disto, esquemas de detecção de erros grosseiros adicionais são exigidos para achar a medição com problemas.

Em 1975, Almasy e Sztano analisaram as propriedades estatísticas das variáveis reconciliadas e propuseram uma série de testes coletivos para detecção de erros grosseiros. Para sistemas com restrições lineares com todas as vazões medidas, eles propuseram o *Teste da Máxima Potência* para detecção de um único erro grosseiro nas medidas quando a variância era conhecida. Este teste foi apresentado como tendo a maior probabilidade de detectar corretamente um erro grosseiro quando somente um erro grosseiro está presente nas medidas, visto que, quando são realizados testes múltiplos, o intervalo de confiança é estendido aumentando assim a região de rejeição do teste estatístico.

Em 1982, Mah e Tamhame propuseram o *Teste da Medição*, um dos métodos mais populares para DEG. Este método baseia-se no teste de hipótese de *Neyman-Pearson*, usando os resíduos entre as medições e os seus valores estimados durante a reconciliação de dados. Assim, o teste da medição trata da principal desvantagem dos testes GT e NT– a incapacidade de definir a localização exata do erro grosseiro. Este teste, no entanto, requer que os dados sejam reconciliados antes que seja feita a detecção de erros.

Em 1985, Madron introduziu o conceito de credibilidade da medida pela atribuição de um valor máximo de magnitude do erro grosseiro.

Em 1987, Narashiman e Mah formularam testes para erros grosseiros usando o conceito de *Razão de Máxima Verossimilhança Generalizada (Generalized Likelihood Ratio test – GLR).* Este teste utiliza um modelo para medição e um modelo para a restrição, e com isto pode-se avaliar a presença de diferentes tipos de erros grosseiros. Eles aplicaram o método a processos em estado estacionário e, ao invés de usar uma estratégia de eliminação seriada, usaram uma estratégia de compensação seriada. A maior contribuição do GLR é que este pode identificar diferentes tipos de erros grosseiros.

Em 1993, Harikumar e Narashiman utilizaram esta idéia e formularam testes para erros grosseiros na presença de limites para os valores estimados. Já em 1995, Tong e Crowe propuseram um teste baseado em Análise de Componentes Principais (PCA).

Quando existe mais de um erro grosseiro presente nos dados, a avaliação do conjunto de dados deve ser modificada, pois os métodos para detecção, identificação e estimação de um único erro grosseiro não são consistentes na presença de mais de um erro grosseiro. (Bagajewicz, 2002). Além disto, surgem novos problemas associados a esta situação, como por exemplo:

- Cancelamento de erros grosseiros
- Espalhamento dos erros grosseiros *(Smearing)* gerando falso alarme

- Presença de mais de um tipo de erro grosseiro
- Problemas associados à filosofia dos métodos de detecção de erros grosseiros (Testar a hipótese de se ter um erro grosseiro por vez, ou mais de um erro por vez)
- Perda de observabilidade, ou problemas com variáveis não observáveis na prática
- Perda de redundância
- Escolha do tamanho do erro grosseiro a ser detectado
- Erros equivalentes
- Sistemas que se degeneram
- Como conseguir descobrir erros grosseiros nas variáveis de interesse

Para isto, foi desenvolvida outra classe de testes de detecção, especificamente construída para lidar com múltiplos erros grosseiros. Estes métodos são, em geral, compostos por uma combinação de métodos já existentes, variam conforme a abordagem do problema (detecção de diferentes tipos de erros grosseiros) e são iterativos. Alguns métodos só conseguem lidar com erros do tipo bias, outros foram feitos para lidar com todos os tipos de erros grosseiros, sendo somente modificado o modelo para medição a ser utilizado.

Quanto à forma de detecção, a maioria utiliza o teste global como detector para o conjunto de medidas e depois é aplicada alguma estratégia para gerar uma lista de candidatos de medidas com erros grosseiros. Após esta lista ser gerada, utilizam-se métodos de estimação para a determinação do tamanho do erro e sua correção. Os métodos podem ser separados em categorias que refletem a maneira com que os erros são detectados e posteriormente compensados. Estas são:

1. Eliminatórios: Eliminam as variáveis suspeitas, uma a uma, até que algum critério seja satisfeito. Exemplo: IMT (Serth and Henan, 1986).

2. Compensatórios: Compensam as variáveis suspeitas, uma a uma, até que algum critério seja satisfeito. Exemplo SSCS (Narashiman e Mah, 1987).

3. Compensatórios simultâneos: Ao invés de realizar a correção, uma a uma, esta é feita simultaneamente para todas as medidas suspeitas. Ex. ICS (Devanathan et al., 2000).

4. Combinatórios: Geram a lista de candidatos de alguma das formas anteriores e depois investigam todas as combinações possíveis de erros grosseiros. Ex. SEGE (Sanchez e Romagnoli, 1999).

5. Combinados: Utilizam estratégias de busca nas medidas e nas restrições associadas, alternadamente até que algum critério seja satisfeito. Ex. MTNT (Yang et al., 1995).

6. Robustos: Métodos que usam outras distribuições para as variáveis que são robustas na presença de erros grosseiros. Assim a detecção é realizada com base nas variáveis que sofreram maior ajuste.

Em 1965, Ripps apresentou uma técnica de detecção de múltiplos erros grosseiros que avaliava a influência no valor qui-quadrado do conjunto de dados antes e após a eliminação de variáveis suspeitas.

Serth and Henan (1986) propuseram sete testes incluindo o *teste da medição iterativo* (*IMT*), o *teste da medição iterativo modificado (MIMT)* e um algoritmo de busca combinatória (SC), onde todas as hipóteses possíveis são testadas e a combinação candidata a erro grosseiro é aquela que gera maior influencia na função objetivo. Neste trabalho ainda foi comparada a performance destes novos métodos com o MT e o NT. Eles concluíram que o melhor seria usar uma combinação de diferentes métodos e então explorar o que há de melhor nestes.

Em 1987, Narashiman e Mah propuseram uma *Estratégia de Compensação Seriada* (*SCS ou ainda SSCS)* que utiliza o teste GLR para avaliar as variáveis suspeitas e compensar os erros grosseiros a cada iteração. Este método apresenta a vantagem de poder tratar de diferentes tipos de erros grosseiros, assim como o seu teste base GLR.

Em 1988, Crowe mostrou que para o caso linear a redução na função objetivo, resultante da eliminação de uma medição é exatamente igual à raiz quadrada do Teste da Medição e para o caso bilinear, esta redução é o limite inferior. Já em 1989, o autor definiu o teste da medição para as restrições remanescentes e para as restrições originais. Novamente para o caso linear a redução na função objetivo pela eliminação de uma restrição ou equivalentemente pela adição de uma vazão não medida é exatamente a raiz quadrada do teste da medição correspondente, comprovando o fato de que o MT é realmente um teste no multiplicador de Lagrange relacionado a esta restrição.

Em 1985, Madron demonstrou que sob a hipótese de que o erro grosseiro de certa magnitude específica, a distribuição do ótimo da função objetivo torna-se uma distribuição qui-quadrada não centrada e determinou o número de desvios padrões que devem ser excedidos pelo erro grosseiro de maneira que este seja determinado com probabilidades de 50% e 90%.

Em 1992, Rollins e Davis propõem a *Técnica de Estimação não-tendenciosa* ("*Unbiased Estimation Technique"* - UBET), onde a maneira com que os testes de hipótese são construídos difere dos métodos até então apresentados. Ao invés do teste de hipótese ser em relação à possibilidade de um erro a cada iteração, e a lista de candidatos a erros grosseiros aumentar, parte-se da hipótese de que todas as variáveis contêm erros grosseiros e a lista de candidatos deve ser diminuída. O nome vem da premissa de que se todos os erros grosseiros presentes forem identificados, a estimação do conjunto de dados será não-tendenciosa.

Em 1994, Sanchez e Romagnoli propuseram uma estratégia combinatorial de *Estimação simultânea de Erros Grosseiros* (*SEGE*) para gerar a lista de variáveis candidatas a conterem erros grosseiros e compensá-las. Uma restrição é submetida ao teste de hipótese por vez e assim é gerada a lista de candidatos. Após, verifica-se a combinação destes candidatos que mais afeta o valor da função objetivo do problema de reconciliação associado. Esta estratégia foi formulada para tratar tanto de erros tipo bias como vazamentos. A escolha do tipo de erro depende do analista.

Keller et al (1994) propuseram uma modificação do algoritmo *SSCS*, apresentado por Narashiman e Mah (1997), transformando o algoritmo de compensação seriada em um algoritmo de compensação simultânea, chamado de *Estratégia Modificada de Compensação Seriada* (*MSCS)*. Assim, somente o tipo e a localização dos erros grosseiros identificados nas etapas anteriores são considerados corretos e as estimativas não são utilizadas na compensação. Este algoritmo também é chamado de CGLR por Bagajewicz e Jiang (1999).

Em 1995, Yang apresentou uma técnica de detecção de erros grosseiros baseada na combinação entre o teste da medição e o teste da restrição chamado teste *MT-NT combinado*. Ele utiliza o teste *MT* para gerar uma lista de candidatos e o *NT* para checar a validade da lista.

Em 1995, Tong e Crowe propuseram a utilização da Análise de Componentes Principais (*PCA*) para a detecção de Erros grosseiros chamada de *Análise da Contribuição*. As vantagens seriam a remoção da correlação entre as variáveis e o controle da probabilidade de o algoritmo cometer erro Tipo I (*falso alarme*). Isto se deve ao fato do teste aplicado ser multivariado, fazendo com que o nível de significância modificado seja menor do que o teste univariado equivalente (Por exemplo, MT de máxima potência). Isto não é necessariamente verdadeiro, pois se pode obter o mesmo efeito escolhendo um nível de significância menor (Narashiman e Jordache, 2000). Além disto, os autores indicaram que o teste tem melhor performance na detecção de erros persistentes com melhor poder de detecção, quando comparado com outros testes. Estas afirmações não foram confirmadas no estudo comparativo realizado por Jordache e Tilton (1999).

Já em 1996, Tong e Crowe propuseram uma estratégia de detecção baseada em *PCA* e *Análise Seqüencial* e tem por objetivo a detecção de erros grosseiros persistentes. Esta usa um teste seqüencial para PCA, chamado *Teste Seqüencial da Razão de Probabilidade (SPRT*) e,

uma vez identificado o erro, é utilizada a metodologia da *Análise da Contribuição* apresentada em 1995.

Em 1996, Rollins et al. propuseram a *Técnica de Combinação Linear* (*LCT*), onde a proposta é avaliar a combinação de restrições que não passam no teste da restrição. Esta estratégia é equivalente a utilizar a técnica de agregação nodal para eliminar nós suspeitos de conterem erros grosseiros e é baseada na técnica *UBET* apresentada em 1992. O método inicia com a lista de variáveis suspeitas contendo todas as variáveis medidas (princípio da técnica *UBET*) e aplica o teste Global para as combinações de nós (restrições) de modo que, a lista de suspeitos é reduzida. A idéia é superar o problema de cancelamento de erros grosseiros nas restrições.

Bagajewicz (1996) aborda o problema da distribuição de probabilidade dos erros na reconciliação de dados. Na quase totalidade dos trabalhos encontrados na literatura disponível, a formulação do problema de reconciliação recai na suposição da distribuição normal dos erros. O autor aponta que esta suposição é verdadeira para dispositivos de leitura que efetuem somente transformações lineares sobre as medidas. No caso de leituras como as de vazão volumétrica, que implicam em transformações não-lineares sobre uma medida primária, mesmo que esta tenha uma distribuição normal, a leitura final a distorcerá e a formulação "clássica" torna-se ineficaz para captar a natureza dos erros nos procedimentos de reconciliação de dados.

Com base no trabalho iniciado em 1996, em 1998, Bagajewicz e Jiang apresentam a teoria dos erros equivalentes. Dois conjuntos de erros são equivalentes quando apresentam o mesmo efeito na reconciliação de dados e são indistinguíveis teoricamente. Assim, quando um conjunto de erros grosseiros é identificado, existe uma possibilidade igual que a localização verdadeira deste conjunto esteja em um dos seus conjuntos equivalentes. Neste mesmo trabalho foi definido o conceito de *degeneração*, onde um conjunto com menor número de erros grosseiros pode ser equivalente ao conjunto verdadeiro.

Em 1999, Sanchez propôs o *MSEGE*, uma modificação do *SEGE,* baseada na teoria dos erros equivalentes proposta por Bagajewicz, onde é incluída a classificação dos conjuntos equivalentes no problema de detecção dos erros grosseiros.

Em continuação ao trabalho apresentado em 1998, Jiang e Bagajewicz propõem uma estratégia de identificação seriada com compensação coletiva para múltiplos erros grosseiros (*SICC*). Nesta estratégia são usados os conceitos introduzidos no trabalho publicado em 1998 em conjunto com o teste Nodal. Neste trabalho também é sugerida uma estratégia para detecção de vazamentos utilizando o MT e a teoria de bias equivalentes e dois algoritmos já existentes (UBET e MSCS) são modificados para levar em conta a teoria dos erros equivalentes. Assim são sendo sugeridos dois novos algoritmos, o *MUBET* e o *MCGLR*.

Em 2000, Devanathan et al. apresentaram uma técnica alternativa para detecção de erros grosseiros chamada de *Estratégia de Correlação dos Balanços* ("*Imbalance Correlation Strategy"* - ICS). Os autores afirmam que esta apresenta alta probabilidade de correta

identificação e baixa probabilidade de cometer erro tipo I. O método é baseado na observação de mudanças na correlação amostral entre os balanços materiais em cada nó e as variáveis associadas a este.

Em 2001, Soderstrom et al. propuseram um método para identificação simultânea de erros grosseiros e reconciliação baseado em *programação mista inteira linear* (MILP) onde o problema de DEG é incluído no problema de reconciliação de dados e as variáveis com bias são detectadas simultaneamente com a estimação dos EGs. Além disto, foi proposta a inclusão dos testes estatísticos populares em DEG nas restrições do problema e o resultado não foi superior aos métodos tradicionais. Como vantagem seria a utilização em problemas grandes e fácil adaptação aos problemas não-lineares.

Zhang *et al.* (2001) mostram que por vezes a eliminação de medidas detectadas contendo erros grosseiros provoca perda de precisão na solução do problema de reconciliação de dados. Desta forma é proposto um método de análise de redundância, baseado na medida da precisão da reconciliação, que permite a eliminação criteriosa e seqüencial de medidas portadoras de erros grosseiros com o objetivo de preservar a solvabilidade do problema de reconciliação.

Além destes métodos ainda podem ser considerados aqui os métodos já apresentados no item 2.2 (Reconciliação de dados), que versam sobre a solução do problema de RD simultânea ao problema de DEG, como por exemplo, *Reconciliação Robusta* e *Error in Variables*. Estes, apesar de não serem métodos de detecção propriamente ditos, podem ser utilizados para tal fim pela utilização dos ajustes realizados na reconciliação como base para os testes estatísticos já apresentados.

Neste trabalho foram implementados e são avaliados os seguintes testes: MT, NT, GLR, PCA, IMT, MIMT, SSCS, SEGE, LCT, MTNT, NTMT. Os testes são avaliados quanto à habilidade de detecção e localização dos erros grosseiros e, posteriormente, são utilizados na estratégia de reconciliação completa, onde a comparação é realizada com métodos de reconciliação robusta para avaliação do resultado do problema de estimação do conjunto de dados propriamente dito.

2.5 Classificação das Variáveis

Em processos industriais de média ou larga escala, existem centenas (por vezes milhares) de variáveis e, por questões técnicas e econômicas, não é possível medir todas elas. Devido a isto, é interessante saber se as variáveis não-medidas podem ser estimadas, ou se em caso de falha de um instrumento, a variável em questão e o conjunto de dados podem ser estimados.

Em reconciliação de dados, Vaclavek (1976) foi o precursor no tratamento desta questão. O autor definiu os conceitos de observabilidade e redundância das variáveis. Crowe et al. (1983) propôs um método que utiliza a decomposição QR para realizá-la. E este é um dos métodos de análise de redundância utilizados neste trabalho e detalhado mais adiante. Na

Figura 2.7 é apresentado um esquema com a classificação das variáveis de processo. As variáveis medidas podem ser classificadas de duas maneiras: (Crowe, 1989)

- ✓ *Variável redundante:* A variável medida é chamada redundante quando ela continua sendo observável mesmo após a sua medição ser removida.

- ✓ *Variável não redundante:* A variável medida é chamada não redundante quando ela é não-observável após a sua medição ser removida.

Da mesma forma, as variáveis não medidas podem ser classificadas em:

- ✓ *Variável observável:* A variável é chamada observável se pode ser estimada a partir das variáveis medidas usando as equações do modelo.

- ✓ *Variável não observável:* A variável é chamada não observável se não pode ser estimada a partir das variáveis medidas usando as equações do modelo.

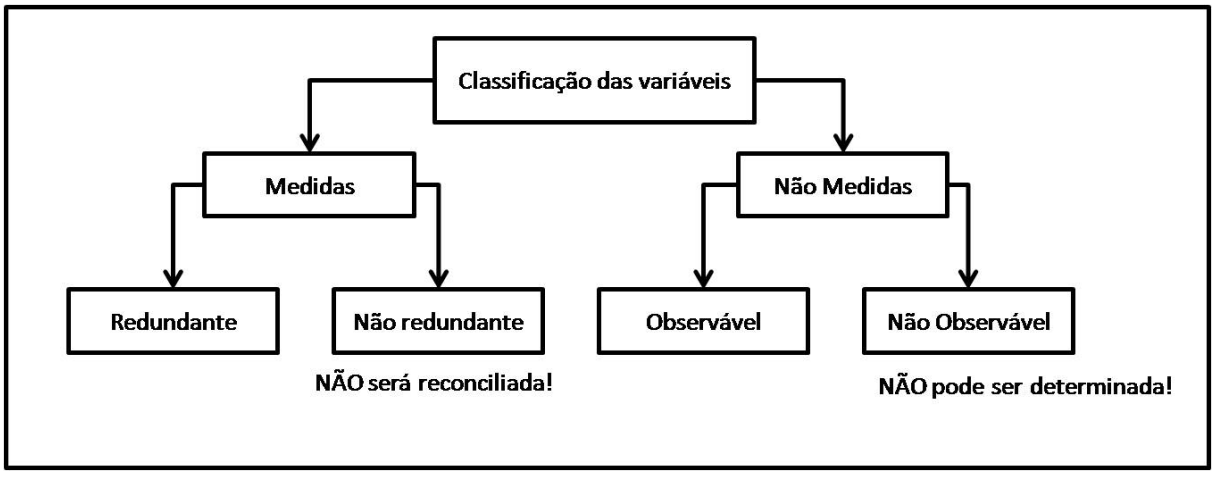

Figura 2.7: Classificação das variáveis de processo.

A classificação é aplicada para reduzir o conjunto das restrições eliminando as variáveis não medidas e as não redundantes e dar indicativos de como formular o problema de RD. É claro que a formulação do problema de otimização de RD e a sua interpretação estatística são fortemente ligadas à unicidade da solução do problema de otimização. Assim é essencial que a classificação seja realizada para que se saiba, de antemão, quais variáveis não redundantes não serão reconciliadas e quais variáveis não medidas não podem ser estimadas.

O objetivo da RD é aumentar a precisão do conjunto de dados medidos explorando somente a propriedade de redundância. A diminuição na redundância de uma variável faz como que a precisão da variável reconciliada diminua. Similarmente, um pré-requisito essencial para a detecção de erros grosseiros é a redundância nas medidas. Teoricamente, é possível identificar um erro grosseiro somente em medidas redundantes. Isto acontece, pois uma medida não redundante é eliminada e não participa do problema reduzido de

reconciliação. Além disto, nenhum teste estatístico pode ser feito com uma medida não redundante.

Existem 2 tipos de redundâncias:

- ✓ *Redundância espacial: utilizada para reconciliação em geral.* Tipicamente, em qualquer processo, as variáveis têm relação, umas com as outras, por restrições físicas como leis de conservação de massa e energia. Dado um conjunto de restrições, um número mínimo de medidas é necessário para calcular todos os parâmetros e variáveis do sistema. Se existem mais medidas que este mínimo, então existe a redundância e esta pode ser explorada. Esta é chamada de *redundância espacial* (Narashiman e Jordache, 2000) e o sistema de equações é dito redundante (sobre determinado). A RD não pode ser feita sem redundância espacial, pois sem nenhuma informação adicional, o sistema é chamado *não redundante* e nenhuma correção nos dados medidos é possível. Caso o sistema apresente menos variáveis medidas que as necessárias, o sistema é chamado de *não determinado* e os valores de algumas variáveis podem ser estimados somente por outros meios ou se medições adicionais forem providenciadas.

- ✓ *Redundância temporal*: As medições de processo são obtidas continuamente no tempo com uma alta taxa de amostragem, produzindo mais dados do que o necessário para determinar o estado estacionário do processo. Se o processo está em estado estacionário, então a redundância temporal pode ser explorada para aumentar a precisão da reconciliação. No caso da RD estacionária, a redundância temporal pode ser utilizada como um pré-tratamento dos dados gerando uma pré-redução de variabilidade na variável a ser reconciliada e diminuindo a presença de *outliers (i.e. grandes picos ocasionais de curta duração, muitas vezes, 1 tempo de amostragem)*.

Além do método já citado, que utiliza a decomposição QR para realizar a classificação das variáveis, ainda existem outros trabalhos que versam sobre este tema na literatura. Madron e Veverka (1992) propuseram um algoritmo de múltiplas eliminações Gauss-Jordan. Pela decomposição de certa matriz e a realização de permutações entre colunas, as variáveis medidas são divididas em redundantes e não redundantes.

Ali e Narashiman (1995) aplicaram a teoria gráfica para analisar a rede de sensores, classificarem as variáveis e medir o grau de redundância. Neste, o *grau de redundância de uma variável* é definido como sendo o número de maneiras que o valor de uma variável pode ser estimado, diretamente ou indiretamente. E *o grau de redundância do sistema* é definido como sendo *o número de variáveis redundantes* (Heyen, Marechal e Kalitventzeff, 1996). Ambas as definições levam em conta a topologia do sistema, mas não consideram a precisão dos sensores.

Além da classificação já apresentada ainda existe uma classificação referente ao comportamento das variáveis durante o procedimento de RD. Na literatura são reportadas dificuldades em trabalhar com variáveis que apresentam características especificas que, mesmo classificadas teoricamente como redundantes e observáveis, na prática se comportam como se não o fossem. Isto ocorre devido à natureza das restrições ou à magnitude dos seus desvios padrões. Estas são chamadas de *variáveis não redundantes na prática e variáveis não observáveis na prática*

Uma *variável não redundante na prática* sofrerá ajustes insignificantes durante a etapa de reconciliação, que corrigirá mais outras variáveis que não continham erros para que as restrições sejam satisfeitas. Mesmo que estas sejam classificadas teoricamente como redundantes, estas não o são e será de difícil identificação a presença de erros grosseiros neste tipo de medida. Uma série de autores reporta dificuldades na reconciliação e detecção deste tipo de variável, entre eles Jordache (1985) e Crowe (1988).

Na literatura encontram-se uma série de características que devem possuir estas variáveis: (Jordache, Mah e Tamhane, 1985; Madron, 1985 e 1992; Crowe, 1988)

I. Variáveis medidas que se relacionam com o conjunto de dados por uma só equação de balanço (principalmente se esta equação tem uma variável não medida)

II. Correntes paralelas

III. Variáveis com pequenos desvios padrões quando comparadas às variáveis pertencentes ao mesmo nó.

Madron e Verveka (1992) definiu que uma *variável é não-redundante na prática* se o seu *coeficiente de ajustabilidade* (a_i) for menor que 0,1. Sendo este definido como:

$$a_i = \left(1 - \frac{\sigma_{\hat{x}_i}}{\sigma_{y_i}} \right) \leq 0.1 \tag{2.11}$$

onde a_i é o coeficiente de ajustabilidade, $\sigma_{\hat{x}_i}$ é o desvio padrão da variável i reconciliada e σ_{y_i} é o desvio padrão da variável medida i. Este coeficiente pode ser utilizado para prever o quanto a precisão da medição i pode ser aumentada pela reconciliação de dados.

Já a classificação das variáveis não observáveis na prática foi proposta por Crowe, 1983. Segundo o autor, *"uma variável não medida é pouco observável se existe uma medida não redundante que, se removida do conjunto, transforma a variável observável em não observável."*.Quando isto acontece, a solução do problema de RD não é única, e assim o problema precisa ser decomposto para que uma solução única seja encontrada. Um exemplo é uma variável não medida observável que a sua estimativa apresenta um desvio padrão tão grande a ponto de ser considerada como variável não observável na pratica. Gomolka et al

(1992), Crowe (1989), Maquin et al (1989) e Ragot et al (1991, 1996) trataram do problema de classificação, usando técnicas gráficas, e estendendo para sistemas bilineares e não lineares. Charpentier (1991) utiliza o índice de detectabilidade, dado por:

$$d_i = \sqrt{\left(I - \sigma^2_{\hat{x}_i} / \sigma^2_{y_i}\right)} \tag{2.12}$$

Quanto maior o fator de detectabilidade, maior é a probabilidade de o erro grosseiro ser identificado e fatores grandes implicam na facilidade de detecção de erros grosseiros menores.

É necessário deixar claro que a falta de redundância e observabilidade só podem ser resolvidas com a adição de medidas adicionais. Isto pode ser feito pela instalação física de outros medidores ou adição de algum modelo que faça o cálculo indireto da variável. Como este problema não tem uma solução óbvia, o foco passa a ser o tratamento de dois outros problemas: influência da perda de redundância no erro da estimação das variáveis reconciliadas e a determinação de variáveis não medidas observáveis.

Capítulo 3

A Reconciliação de dados

3.1 Formulação Geral do Problema de RD

O objetivo da Reconciliação de Dados (RD) é obter estimativas mais precisas para medições de variáveis de processo, utilizando a condição com que estas se relacionam, ou seja, um modelo matemático. Assim, o problema geral de reconciliação é definido formalmente como: *"Dado um conjunto de variáveis medidas $\{y_1,...,y_m\}$ referentes a um conjunto de variáveis do modelo $\{x_1,...,x_m\}$, deseja-se obter as melhores estimativas para estas variáveis medidas, $\hat{x}$, que satisfaçam as relações essenciais dadas pelo modelo do processo."* (adaptado de Bagajewicz, 2000).

Desta forma, o problema de otimização é definido *tradicionalmente* como um problema de minimização do erro quadrático entre as variáveis medidas e as do modelo, sujeitas às restrições, do tipo Mínimos Quadrados Ponderados,

$$\begin{aligned} &\min_{x} \left\{ S = (y-x)^T W^{-1} (y-x) \right\} \\ &\text{sujeito à} \\ &h(x,p,t) = 0 \\ &g(x,p,t) \le 0 \end{aligned} \tag{3.1}$$

onde W é uma matriz de ponderação, h é um conjunto de equações de igualdade que corresponde ao modelo matemático do processo, g é um conjunto de equações de desigualdade representando limites operacionais e de validade, os quais as estimativas devem satisfazer. Estes relacionam as variáveis do modelo (x), parâmetros (p) e tempo (t). Substituindo o modelo da medição, onde se supõe somente a presença de erros aleatórios e já apresentado na equação 2.3, tem-se:

$$\begin{aligned} &\min_{\varepsilon} \ \varepsilon^T W^{-1} \varepsilon \\ &\text{sujeito à} \\ &y = x + \varepsilon \\ &h(x, p, t) = 0 \\ &g(x, p, t) \le 0 \end{aligned} \tag{3.2}$$

Na formulação geral, o modelo do processo é dado por um conjunto de equações algébrico-diferenciais, representando as leis de conservação de massa, energia, mas também podem ser funções quaisquer que relacionam as variáveis. A motivação para a utilização das leis de conservação é que estas são essenciais para qualquer sistema e existem poucas chances de existirem falhas na sua estrutura (Crowe, 1996), mesmo que ainda possam existir erros em parâmetros. Este conjunto de restrições pode ser composto de balanços lineares (conservação de massa), balanços por componentes (bi lineares), balanços de energia (não lineares) e estes podem ser dependentes do tempo, das variáveis de estado e de parâmetros (determinísticos ou aleatórios). Assim, o tipo de restrição define a técnica utilizada para a solução do problema, com o já mencionado no capítulo 2.

O foco deste trabalho são sistemas lineares e em estado estacionário. Esta escolha foi feita porque existe uma grande variedade de técnicas para a solução do problema conjunto de reconciliação de dados e detecção de erros grosseiros aplicadas a estes sistemas. Além disto, apesar desta dissertação não tratar dados reais de processo, esta escolha faz com os casos simulados estejam bastante próximos da realidade.

O interesse em testar uma grande variedade de técnicas relacionadas com estes sistemas é porque estas são as bases para a solução de problemas mais complexos, onde nem sempre o grande desafio é o problema de reconciliação em si. Por exemplo, em processos não-lineares, o grande desafio são os modelos (assim como em sistemas dinâmicos). Neste caso, a abordagem da reconciliação propriamente dita é muito simplificada, pois não se tem relações óbvias para transformações não-lineares das funções distribuição propostas para o tratamento dos erros. Isto faz com que a inclusão da incerteza das variáveis e dos parâmetros seja mais difícil de ser tratada (que é o objetivo da reconciliação!) e muitas vezes nem faça sentido ser adicionada ao problema, pois o modelo já apresenta incertezas.

Outra motivação é formar uma base teórica e uma visão crítica dos métodos para, no futuro, trabalhar na extensão das técnicas existentes para estado estacionário, para lidar com estes sistemas mais complexos mencionados acima. Assim, partir-se-á da formulação geral e no próximo item, as premissas serão apresentadas e a formulação será simplificada.

3.1.1 Reconciliação de Dados Linear para Sistemas em Estado Estacionário

Partindo da formulação geral, ao considerar que o processo está em estado estacionário, equivale a dizer que a média das variáveis ao longo do tempo, assim como a sua variância são constantes (Arora e Biegler, 2001). E as restrições do problema, se o modelo do

processo é linear, resumem-se aos balanços de massa e qualquer outra relação linear entre as observações, as variáveis do modelo e qualquer outro parâmetro desejado. Assim, as restrições do problema de otimização geral de reconciliação de dados (Equação 3.2) podem ser simplificadas para,

$$\begin{aligned} &Ax = 0 \\ &A \in \Re^{m,n},\ x \in \Re^{m} \end{aligned} \tag{3.3}$$

Ou ainda

$$Ax = c \tag{3.4}$$

onde x corresponde ao vetor de m variáveis do modelo do processo, A é uma matriz n x m *também chamada de matriz de incidência*. No caso da reconciliação estacionária linear, a matriz de incidência é formada pelas equações do balanço de massa onde seus elementos são iguais a 1, -1 ou 0, dependendo se a corrente associada à restrição é de entrada, saída ou não participa do balanço. Caso alguma variável seja exatamente conhecida, a equação pode ser igualada a um valor c constante como mostra a equação 3.4. Assim, o problema de reconciliação de dados linear em estado estacionário pode ser expresso por:

$$\begin{aligned} &\min_{\varepsilon} \varepsilon^T W^{-1} \varepsilon \\ &sujeito\ à \\ &y = x + \varepsilon \\ &Ax = 0 \end{aligned} \tag{3.5}$$

onde W é uma matriz quadrada, m x m, de ponderação. Este problema pode ser resolvido por diferentes técnicas, sendo que a mais utilizada é a solução analítica do problema de otimização sem restrições obtida pelo o *Método dos Multiplicadores de Lagrange*.

3.1.2 Obtenção da Solução Analítica do problema de RD sem restrições - Método dos Multiplicadores de Lagrange

O método dos multiplicadores de Lagrange transforma o problema de otimização com restrições em um problema sem restrições. Para isto, se expressa o problema de otimização como:

$$\max_{\lambda} \min_{\varepsilon} \{ L(\lambda, y) = (\varepsilon W^{-1} \varepsilon) - \lambda^T (A\varepsilon + Ay) \} \tag{3.6}$$

onde λ são os multiplicadores de Lagrange. Tomando as derivadas em relação à ε e λ tem-se as equações 3.7 e 3.8.

$$\frac{dL}{d\varepsilon} = 2W\varepsilon - A^T \lambda \tag{3.7}$$

$$\frac{dL}{d\lambda} = Ay - A\varepsilon \tag{3.8}$$

Igualando estas derivadas a zero, a fim de determinar o ponto ótimo, é possível obter os valores para ε a partir da Equação 3.7:

$$\varepsilon = \frac{1}{2} W^{-1} A^T \lambda \tag{3.9}$$

Aplicando este na Equação 3.8, é possível obter-se também uma relação para λ representada pela Equação 3.11.

$$0 = -\left(\frac{1}{2} A W^{-1} A^T \lambda - Ay \right) \tag{3.10}$$

$$\lambda = -2 A W^{-1} A^T (Ay) \tag{3.11}$$

Substituindo a Equação 3.11 no valor de ε, obtém-se a solução analítica do problema de otimização:

$$\hat{x} = y - W^{-1} A^T (A W^{-1} A^T)^{-1} (Ay) \tag{3.12}$$

onde $\hat{x}$ são as estimativas dos valores reais das variáveis de interesse ou também chamados de *valores reconciliados* e satisfazem as restrições representadas pela matriz *A*.

A solução analítica apresentada tem a desvantagem de poder resultar em valores negativos que muitas vezes não tem significado físico, visto que não existe a possibilidade de adicionar limites para as variáveis. Caso seja necessário adicionar limites, podem-se utilizar métodos de programação quadrática para a solução do problema de reconciliação e este será apresentado mais adiante ainda neste capítulo.

3.1.3 Abordagem Tradicional do Problema de RD – Formulação Estatística e suas Conseqüências

O problema de reconciliação de dados é tratado tradicionalmente usando uma base teórica estatística, o que pode não só ajudar no entendimento desta técnica, como também adicionar informações úteis sobre como aumentar a acuracidade dos dados obtidos e as propriedades estatísticas das estimativas resultantes (Narashiman e Jordache, 2000).

Se objetivo principal da RD é reduzir a variância das variáveis de processo (aumentar a precisão, fechar os balanços de massa – todos considerados sinônimos neste caso) isto é o mesmo que tratar especificamente o erro aleatório presente nos dados. Para isto, já foi apresentada a modelagem deste e as premissas propostas para seu tratamento. O natural é utilizar estas premissas, refinando o problema de otimização de mínimos quadrados.

Assim, observando a função objetivo proposta na Equação 3.5, pode-se concluir que a simples minimização do erro entre o valor reconciliado (estado estimado) e a medição (estado observado), garante somente que as estimativas são ótimas do ponto de vista que estarão próximas deste estado observado. Mas isto não garante a obtenção de estimativas que reflitam o estado verdadeiro (e desconhecido). Desta maneira, a matriz *W* (de ponderação da função objetivo) tem um papel importante na solução do problema de otimização, a sua escolha pode ser feita de maneira a melhorar a estimação.

O argumento para esta escolha vem da *desigualdade de Chebyshev* que diz que: *"Se o desvio padrão de uma variável aleatória for utilizado como unidade de medida, a probabilidade da estimativa de uma variável estar longe do seu valor esperado (μ) é pequena. Mais precisamente, para qualquer variável aleatória X que possua um desvio padrão σ, a probabilidade da estimativa de X estar no mínimo kσ vezes afastada do seu valor esperado μ não pode ser maior que* $1/k^2$*"* (Beck, 1977). A seguir são mostradas duas formas de expressar a desigualdade de Chebyshev:

$$P\left(\left|X-\mu\right| \geq k\,\sigma\right) \leq \frac{1}{k^2} \tag{3.13}$$

$$P\left(\left|X-\mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} \tag{3.14}$$

A simples minimização determinística do erro não reflete o comportamento aleatório deste. Para isto, esta informação pode ser incluída no problema de otimização de modo que a matriz *W* tenha como elementos da diagonal a variância σ_i^2 da medida *i* e os elementos fora da diagonal são dados pela covariância dos erros entre as variáveis *i* e *j*, σ_{ij}^2. Desta forma, este caráter aleatório do problema passa a ser incluso no problema de otimização e esta matriz *W* também é conhecida como *matriz de pesos estatísticos* (Wolberg, 2005).

Além disto, define-se que as observações (medidas) são consideradas independentes e conseqüentemente não existe correlação entre as variáveis, tornando assim *W* uma matriz diagonal. Como conseqüência, o problema de otimização se reduz a,

$$\begin{aligned} &\underset{x}{Min}\ \ S = \left\{ \sum_{i=1}^{n} \frac{\left(y_i - x_i\right)^2}{\sigma_i^2} \right\} \\ &sujeito\ à \\ &A.x = 0 \end{aligned} \tag{3.15}$$

Portanto, a ponderação é feita de modo que a variável menos confiável (maior desvio padrão) tenha um ajuste maior que outra variável com desvio padrão menor. Além disto, sendo o erro aleatório gaussiano, identicamente distribuído e independente, os valores estimados estarão tão próximos probabilisticamente do estado verdadeiro desejado, quanto à desigualdade de Chebyshev permitir.

3.1.4 Estimação de Máxima Verossimilhança (MLE)

Voltando à formulação geral do problema de reconciliação de dados, a construção do problema de otimização na abordagem clássica é feita de maneira que o estimador obtido seja condizente com as propriedades das variáveis a serem estimadas. Entretanto, ao invés de procurar algum estimador que tenha a mesma forma da variável observada (variância da medição), e utilizar o estimador que tenha o mesmo significado (mínimo de uma função quadrática), pode-se utilizar um método mais geral para a obtenção deste. Um destes métodos é o da Máxima Verossimilhança. Este procura escolher entre os possíveis valores do estado estimado, aquele que maximiza a probabilidade de se obter aquela observação (Beck e Arnold, 1974).

Deste modo, derivando o problema de reconciliação de dados como um problema de Estimação de Máxima Verossimilhança (MLE) onde, dado um conjunto de medidas, a probabilidade das variáveis estimadas (reconciliadas) é maximizada como mostra a Equação (3.16).

$$\max_{x} p(x|y) \tag{3.16}$$

De acordo com o *Teorema de Bayes*, a probabilidade das variáveis do modelo do processo dada as medidas, pode ser escrita em termos da probabilidade das medições *p{y|x}* e da distribuição de probabilidade das variáveis do modelo *p{x}*.

$$\max_{x} \frac{p(y|x)p(x)}{p(y)} = \max_{x} p(x|y) \tag{3.17}$$

O denominador da Equação (3.17) é independente das variáveis do modelo, não afeta o problema de otimização e somente atua como uma constante de normalização. O primeiro termo no numerador representa a densidade de probabilidade das medidas com base nas variáveis do modelo, a qual é a distribuição dos erros de medição, *p(y-x)*. Se as *i* medidas são consideradas independentes, o termo *p(y|x)* é calculado como:

$$p(y|x) = \prod_{i} p(y_i|x_i) \tag{3.18}$$

O termo $p(x)$ é uma variável binária. Esta assume valor igual a 1 caso as restrições sejam satisfeitas (sob a hipótese de que o termo $p(x)$ é convertido no conjunto das restrições, o problema original é convertido em um problema de otimização com restrições). Caso contrário assume valor igual à zero.

Para comparar com o estimador escolhido na abordagem tradicional pode-se assumir que as funções densidade de probabilidade das variáveis reconciliadas e das variáveis de processo são gaussianas. Assim, têm-se as seguintes relações:

$$P(y - x) = P(\varepsilon) \sim N(0, W) \tag{3.19}$$

$$p(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}} \tag{3.20}$$

Se as medições são independentes, os erros nas medidas são independentes e o produto desta probabilidade sobre todos os medidores é definido com:

$$p(x|y) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{x_i - y_i}{\sigma_i}\right)^2\right\} \tag{3.21}$$

Tomando-se o negativo do logaritmo da função objetivo de maximização representada pela Equação (3.21) resulta no problema de minimização convencional de mínimos quadrados ponderados. Se a matriz *W* for diagonal, a forma resultante é,

$$\underset{x}{Min} \;\; S = \left\{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2}\right\} \tag{3.22}$$

Voltando a solução analítica do problema de mínimos quadrados ponderados, representada pela equação 3.12, pode-se ver que as estimativas são obtidas usando transformações lineares das medidas, ou seja,

$$\hat{x} = y - W^{-1}A^T(AW^{-1}A^T)^{-1}Ay = [I - W^{-1}A^T(AW^{-1}A^T)^{-1}A] = Cy \tag{3.23}$$

Estas também serão normalmente distribuídas, com valor esperado:

$$E[\hat{x}] = CE[y] = Cx = x \tag{3.24}$$

O que demonstra uma propriedade conhecida de MLE de que as estimativas são não-tendenciosas e com covariância dada por:

$$Cov[\hat{x}] = E[(Cy)(Cy)^T] = CWC^T \tag{3.25}$$

O maior interesse na abordagem do problema de reconciliação como um problema de máxima verossimilhança é mostrar que a técnica possibilita a utilização de qualquer outro tipo de função distribuição do erro nas variáveis medidas e a obtenção de outros tipos de funções objetivo para o problema de reconciliação.

No caso da distribuição Gaussiana, o máximo da função de máxima verossimilhança e o mínimo da solução de mínimos quadrados é a média do erro aleatório, e a solução dos dois problemas coincide. A escolha da matriz de ponderação como sendo *W* apresenta uma série de vantagens. Desta forma garante-se que o estimador provê estimativas ótimas, não-tendenciosas, de variância mínima, e a probabilidade de se obter as observações é máxima. Se

for interessante obter as provas matemáticas, sugere-se Beck e Arnold (1977) ou Jazwinski (1970).

Outro aspecto interessante é que o *teorema de Bayes* requer que a equação (3.17) seja satisfeita, ou seja, os valores reconciliados não podem estar fora dos limites impostos pelos valores reais (o desvio padrão do valor definido na matriz *W*) (Crowe, 1996). Desta forma, a formulação não pode ser utilizada para tratamento de dados com erros grosseiros, validando as afirmações já feitas nos capítulos introdutórios. Para lidar com a presença de erros grosseiros nos dados existem diferentes técnicas. Nesta dissertação foram escolhidas duas técnicas diferentes por parecem as mais promissoras. A mais utilizada na literatura, e que será explorada mais adiante, é o uso de algoritmos de detecção e eliminação de erros grosseiros baseada nas propriedades estatísticas dos erros aleatórios. A segunda técnica, e que é interessante ser introduzida neste momento, é baseada na escolha de outras funções objetivo para o problema de otimização. Com isto pretende-se obter estimativas que não sejam influenciadas pela presença dos erros grosseiros. Esta técnica é conhecida como *Reconciliação Robusta* e é apresentada a seguir.

3.2 Reconciliação de Dados Robusta

Na abordagem tradicional de reconciliação de dados, considera-se que os erros de medição seguem uma distribuição normal com média zero e variância conhecida e todas as inferências estatísticas são baseadas nesta distribuição. Entretanto, se desvios desta distribuição ideal ocorrem, outra função objetivo (que não requer que esta premissa seja verdadeira) pode ser uma melhor candidata.

No caso da distribuição normal, por ser uma distribuição com ponto de *breakdown* próximo a zero, a presença e qualquer erro grosseiro, mesmo que pequeno, invalida a base estatística do estimador e gera estimativas contaminadas (efeito *smearing)*. O "*breakdown point"* é uma medida de robustez do estimador, definida por Gnadesikan (1997) como a maior fração das observações em uma amostra, que podem ter valores extremos sem distorcer o valor do estimador. Existem diversas famílias de funções objetivo que podem ser utilizadas para resolver o problema de reconciliação na presença de erros grosseiros. Funções Objetivo com pontos de *breakdown* mais adequados (por exemplo, distribuições com caudas mais pesadas ou misturadas) que podem ser utilizadas de maneira que os erros grosseiros não sejam espalhados por todo o conjunto de dados.

É justamente deste problema que trata a estatística robusta (Maronna et al, 2006). Pode-se, ao invés de considerar uma distribuição ideal, construir um estimador de maneira que gere resultados sem *bias* na presença desta distribuição ideal e que minimize a sensibilidade em relação aos desvios da idealidade até certo grau (Albuquerque e Biegler, 1996). Isto faz com que as estimativas não sejam contaminadas (ou pelo menos a contaminação seja diminuída). Do ponto de vista probabilístico, podem-se utilizar leis de probabilidade e o principio da máxima verossimilhança, maximizando a função probabilidade do erro de medição e obtendo a função objetivo que satisfaça o princípio da máxima verossimilhança,

$$max\ p(\varepsilon) = max \prod_i p_i \tag{3.26}$$

Esta ainda pode ser generalizada, para uma família de funções objetivo de acordo com a *função objetivo generalizada de máxima verossimilhança* proposta por Huber (1981):

$$min \left\{ J = \sum_{i=1}^{v} \rho(\varepsilon_i) \right\} \tag{3.27}$$

onde ε é o erro padrão

$$\varepsilon_i = \left(\frac{y_i - x_i}{\sigma_i} \right) \tag{3.28}$$

Assim, uma função monótona ρ pode ser usada para a formulação da reconciliação de dados de modo que os erros grosseiros tenham um efeito reduzido na estimação das variáveis medidas. A obtenção do estimador é dada pela escolha da função objetivo ρ e esta não precisa ser uma função quadrática. Para que a estimação seja robusta, esta função objetivo deve ser menos influenciada por valores grandes de ε que a estimação convencional por mínimos quadrados. Nesta dissertação foram testados e comparados três diferentes estimadores: a função *normal contaminada*, função *fair* e a função de *cauchy* e que serão apresentados mais adiante, separadamente.

As funções objetivo utilizadas neste trabalho foram escolhidas pela possibilidade de comparação com trabalhos existentes na literatura, pois a utilização de estatística robusta, na área de reconciliação de dados é relativamente recente (Tjoa e Biegler, 1991; Johnston e Kramer, 1995; Zhang et al, 1995; Albuquerque e Biegler, 1996; Chen et al, 1998; Arora e Biegler, 2001; Wang e Romagnoli, 2003; Özyurt e Pike, 2004; Ragot, 2005; Zhou et al; 2006). Para comparar M-estimadores, do ponto de vista de robustez, se utiliza o conceito de Função de Influência (IF), definida por:

$$IF = \frac{\partial J}{\partial \varepsilon} \tag{3.29}$$

onde J é a função objetivo e ε é o erro de medição. A IF representa efeito de uma medição nas estimativas obtidas. Hampel (1986) define formalmente a IF como sendo uma função que descreve o efeito de uma contaminação infinitesimal no ponto x, referente à estimativa (sensibilidade da função máxima verossimilhança em relação ao erro). Na Figura 3.1 são mostradas as diferentes funções objetivo e suas respectivas funções de influência, onde o termo a*juste* refere-se ao erro padrão (Equação 3.28), e o termo *erro* refere-se ao desvio entre o valor medido e o valor real. As figuras subseqüentes seguem este mesmo padrão.

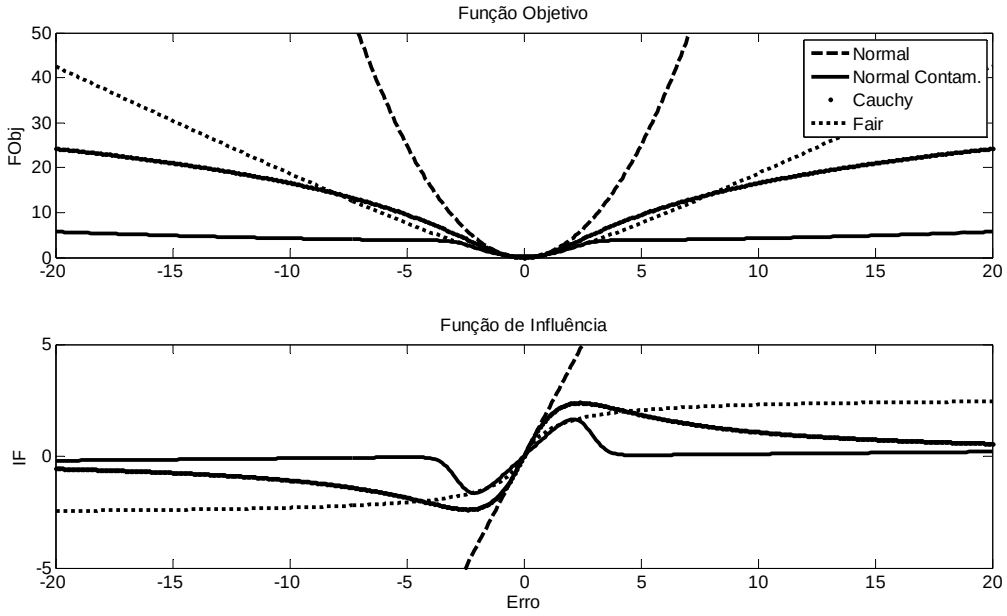

Figura 3.1: Comparação entre as diferentes funções objetivo e respectivas funções de influência.

Pode-se verificar que todas as funções se aproximam da distribuição normal em um pequeno intervalo ao redor de zero e após, cada uma apresenta um comportamento diferente em relação a grandes desvios. O ajuste realizado utilizando a distribuição normal, à medida que o erro cresce, também cresce ilimitadamente. Entretanto as *funções robustas* apresentam um ajuste menor para valores grandes de erro. As funções Normal Contaminada e de Cauchy aproximam-se de zero para valores grandes (a segunda mais lentamente). Já a função *fair* é limitada por um valor constante. Nas seções seguintes, as distribuições serão apresentadas formalmente. A função Gaussiana só será citada novamente para comparação e demonstração da maneira com que são obtidas as funções objetivo e suas respectivas funções de influência.

3.2.1 Distribuição Normal

Como já detalhado anteriormente a função objetivo quadrática é derivada da premissa de que os erros nas medidas seguem uma distribuição normal padrão. A função densidade de probabilidade Gaussiana de um único erro de medição $\varepsilon \sim N(0, \sigma)$ é dada por:

$$p(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2} \right) \tag{3.30}$$

E a sua estimativa de máxima verossimilhança é obtida maximizando-se:

$$p(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m | \mu_i, \sigma_i) = \prod_{i=1}^{m} \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2} \right) \tag{3.31}$$

Sujeito às restrições de igualdade do modelo do processo. A minimização do logaritmo natural da função de máxima verossimilhança apresenta o mesmo mínimo e simplifica o cálculo das estimativas. Os valores constantes não influenciam na solução, estes são removidos da função objetivo.

$$\psi(\varepsilon) = ln\left(\frac{1}{p(\varepsilon)}\right) = \sum_{i=1}^{m} \frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2} - \sum_{i=1}^{m} ln\,\sigma_i - m\,ln\sqrt{2\pi} \tag{3.32}$$

$$J = \sum_{i=1}^{m} \frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2} = \varepsilon^T W^{-1}\varepsilon \tag{3.33}$$

A sensibilidade do erro ou função de influência é dada por:

$$IF = \frac{\partial J}{\partial \varepsilon} = 2\sum_{i=1}^{m_y} \frac{1}{2}\frac{\varepsilon_i}{\sigma_i} = 2W^{-1}\varepsilon \tag{3.34}$$

Na Figura 3.2, a função objetivo e a função de influência para a distribuição Gaussiana é mostrada. A IF, neste caso, aumenta linearmente com o tamanho do erro o que comprova que, na presença de erros grosseiros, esta função objetivo não é robusta e não tem a habilidade de ignorar a contribuição de dados extremos.

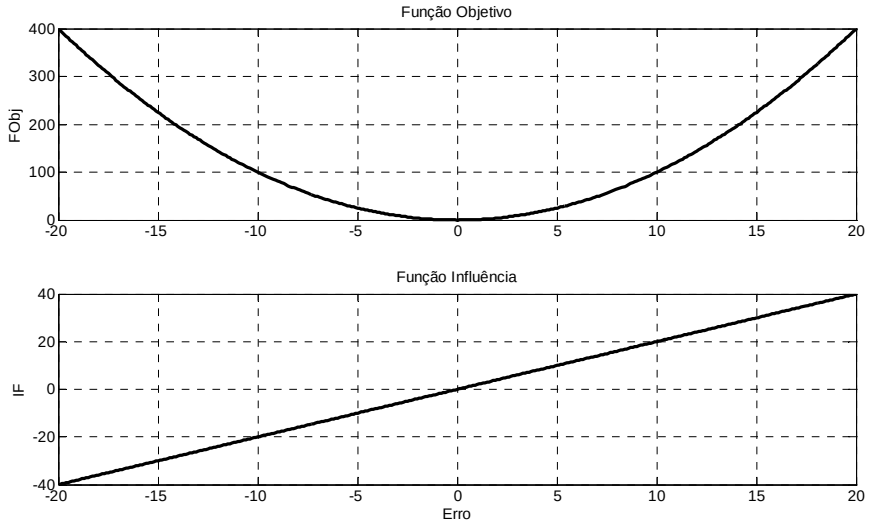

Figura 3.2: Função Objetivo e Função de Influência para distribuição normal.

3.2.2 Distribuição Normal Contaminada

Na presença de erros grosseiros, a combinação linear de duas funções de distribuição normais, baseadas na suas máximas verossimilhanças, é usada (Tjoa e Biegler, 1991). Pode-se partir do princípio de que um percentual *(1-p)* das observações das variáveis é exatamente descritas pela distribuição normal (para fins ilustrativos, chamada *G e* $G \sim N(\mu, \sigma^2)$), e que o restante seja proveniente de uma outra distribuição não conhecida *(H)*, então se pode considerar que a distribuição resultante *(F)* seja,

$$F = (1-p)G + pH \tag{3.35}$$

Em geral, diz-se que *F* é uma mistura (ou combinação) de *G* e *H*, sendo que *H* pode ser qualquer distribuição, por exemplo, uma distribuição normal com variância maior ou média diferente. Quando *G* e *H* são gaussianas, a função *F* resultante é conhecida como *distribuição normal contaminada* ("*contaminated normal"* ou ainda *"normal mixture"*, Maronna, Martin e Yohai, 2006).

Para uma probabilidade de erro grosseiro *p (p<0.5)* e a razão dos desvios padrão dos erros grosseiros e dos erros aleatórios *b (b>1)*, a função para a distribuição gaussiana combinada de um único erro de medida é escrita como sendo:

$$p_{cg}(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}}\left[(1-p)\exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}\right) + \frac{p}{b}\exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2 b^2}\right)\right] \tag{3.36}$$

A estimativa de máxima verossimilhança pode ser obtida maximizando-se:

$$p_{CG}\left(\varepsilon_1, \varepsilon_2, ..., \varepsilon_{m_y} \middle| \mu_i, \sigma_i\right) = \prod_{i=1}^{m_y} \frac{1}{\sigma_i\sqrt{2\pi}}\left[(1-p)\exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2}\right) + \frac{p}{b}\exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2 b^2}\right)\right] \tag{3.37}$$

sujeita às restrições de igualdade do modelo do processo. Isto é o mesmo que minimizar o logaritmo natural da função de verossimilhança, gerando a função objetivo,

$$J_{CG}(\varepsilon) = ln\left(\frac{1}{p_{CG}(\varepsilon)}\right) = -\sum_{i=1}^{m_y} ln\left[(1-p)\exp\left(-\frac{\varepsilon_i^2}{\sigma_i^2}\right) + \frac{p}{b}\exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2 b^2}\right)\right] \tag{3.38}$$

E a função de influência:

$$IF = \frac{\partial J}{\partial \varepsilon} = \frac{\varepsilon}{\sigma^2}\left[1 + \left(\frac{p\left(b^2-1\right)\exp\left(\frac{\varepsilon^2}{2\sigma^2}\right)}{b^2\left(b(p-1)\exp\left(\frac{\varepsilon^2}{2b^2\sigma^2}\right) - p\exp\left(\frac{\varepsilon^2}{2\sigma^2}\right)\right)}\right)\right] \tag{3.39}$$

Assim, cada erro de medida pode ser testado contra a distribuição combinada. Se a probabilidade associada à um erro é maior do que a probabilidade do erro aleatório, então o erro de medida é identificado como erro grosseiro. Deste modo, a distribuição pode ser usada como um teste de detecção de erros grosseiros. Caso as expressões dadas pela Equação 3.40 sejam verdadeiras, então um erro grosseiro está presente na medida *i*.

$$\frac{p}{b}exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2 b^2}\right) > (1-p)exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}\right) \quad ou \quad |\varepsilon_i| > \sigma_i\sqrt{\frac{2b^2}{b^2-1}ln\left(\frac{b(1-p)}{p}\right)} \tag{3.40}$$

Nesta dissertação é utilizado um teste diferente deste. Isto acontece porque não é possível derivar expressões semelhantes para todas as outras funções objetivo. Por este motivo opta-se por um teste mais geral que se baseia na premissa de que, se somente erros aleatórios estão presentes, as funções objetivo robustas realizarão ajuste semelhante ao ajuste realizado pela função objetivo convencional. Na Figura 3.3 são mostradas a função objetivo e a IF para a função gaussiana combinada, comparadas com a distribuição normal. Nesta pode-se verificar que as duas funções objetivo são semelhantes para um intervalo restrito ao redor da média.

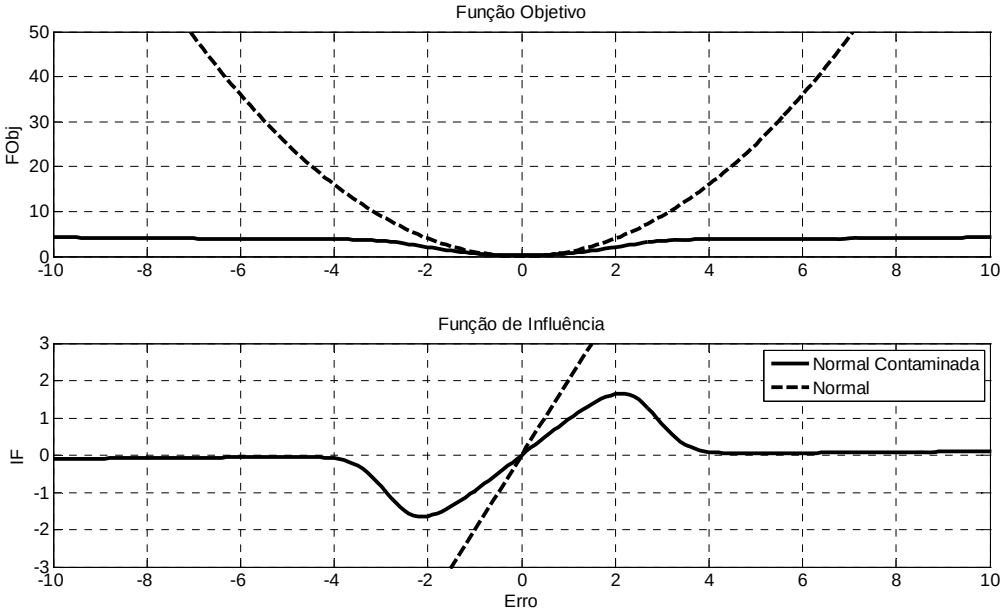

Figura 3.3: Função Objetivo Normal Combinada e Função de Influência (*b=10; p=0,235*).

Nas Figuras 3.4 e 3.5 são ilustradas a influência dos parâmetros *p* (probabilidade de existir erro grosseiro) e *b* (a relação entre o tamanho do erro grosseiro e o desvio padrão do erro aleatório da variável), respectivamente.

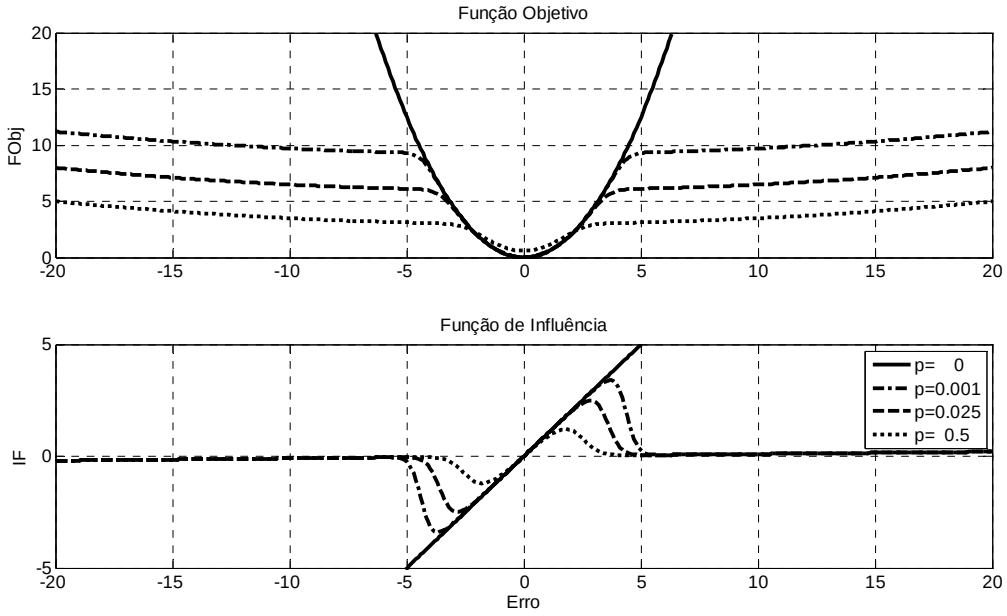

Figura 3.4: Influência do parâmetro *p* na função objetivo CN e na IF.

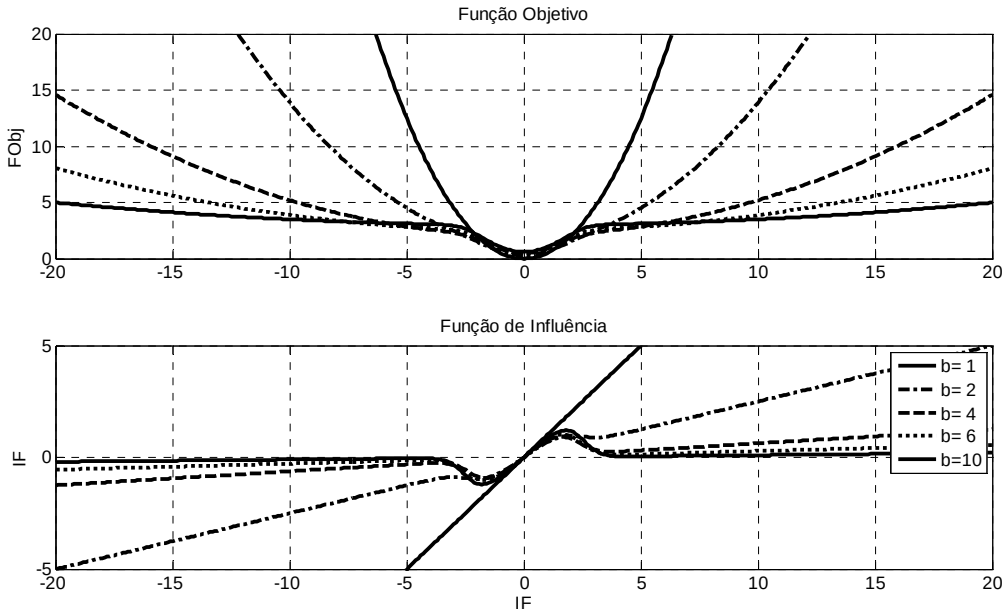

Figura 3.1: Influência do parâmetro *b* na Função Objetivo CN e na IF.

Nas figuras 3.4 e 3.5 fica evidente que a Função Normal Combinada é limitada para grandes valores de erros grosseiros, assim como sua função de influência. Esta restrição pode ser obtida pela mudança dos valores de *b* e *p*, sendo que em geral a sua determinação é tratada com sintonia do estimador. Quanto menor a relação entre o tamanho do erro grosseiro e o desvio padrão do erro aleatório, mais próximo da distribuição normal será o comportamento do estimador, assim como quanto menor for a probabilidade de existência do erro grosseiro.

3.2.3 Função de Cauchy

A função densidade de probabilidade de Cauchy é uma distribuição do tipo cauda-pesada ("*heavy-tailed"* ou *"fat-tailed"*), onde a densidade da cauda tende a zero mais lentamente que a densidade da cauda da distribuição normal. Seu formato é parecido ao da distribuição normal, mas apresenta uma particularidade interessante – a sua média não existe. Esta é um caso particular da família de distribuições t-student, e é dada por:

$$f_\nu(\varepsilon) = c_\nu \left(1 + \frac{\varepsilon^2}{\nu} \right)^{-(\nu+1)/2} \tag{3.41}$$

$$C_\nu = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \tag{3.42}$$

onde Γ é a função Gama. Esta família contém todos os graus de pesos para as caudas ("*heavy-tailedness"*). Quando $\nu \to \infty$, a função tende à densidade normal padrão e quando $\nu = 1$, esta leva o nome de *Distribuição de Cauchy*. A formulação adaptada para reconciliação de dados:

$$p_C(\varepsilon) = \frac{1}{\pi\gamma\left[1 + \frac{\varepsilon^2}{\gamma^2} \right]} \tag{3.43}$$

onde γ é o parâmetro *"half width at half maximun"*. A função objetivo é dada por:

$$J_c = m_y \, ln(\pi\gamma_i) + \sum_{i=1}^{m_y} ln\left(1 + \frac{\varepsilon_i^2}{\gamma_i^2} \right) \tag{3.44}$$

Removendo os valores fixos e tomando $\gamma_i = \sigma_i$,

$$J_c = \sum_{i=1}^{m_y} ln\left(1 + \frac{\varepsilon_i^2}{\sigma_i^2} \right) \tag{3.45}$$

A função de influência da distribuição de Cauchy (IF_c) é dada por,

$$IF_c = \frac{\partial J_c}{\partial \varepsilon} = \frac{2}{\sigma^2} \frac{\varepsilon}{\left(1 + \frac{\varepsilon^2}{\sigma^2} \right)} \tag{3.46}$$

As funções objetivo e as respectivas IFs para a função de Cauchy e a Distribuição Normal são mostradas na Figura 3.6. Nesta fica evidente que função de influência da Função de Cauchy aproxima-se de zero para grandes valores, mas bem mais lentamente que a

distribuição Normal Contaminada. Já na Figura 3.7 é mostrada a influência do parâmetro *c* na distribuição.

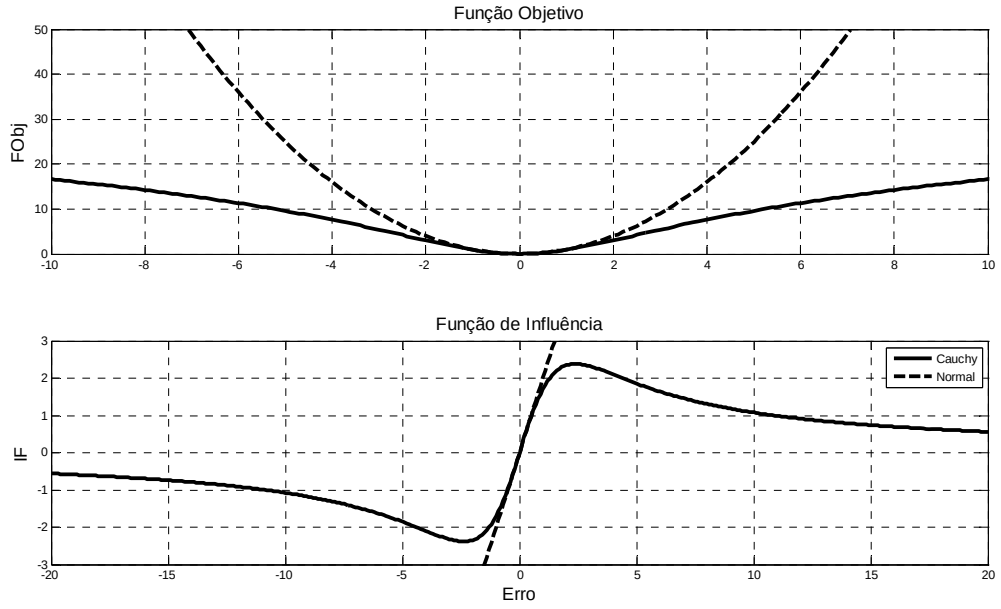

Figura 3.2: Comparação entre a distribuição normal e a função de Cauchy.

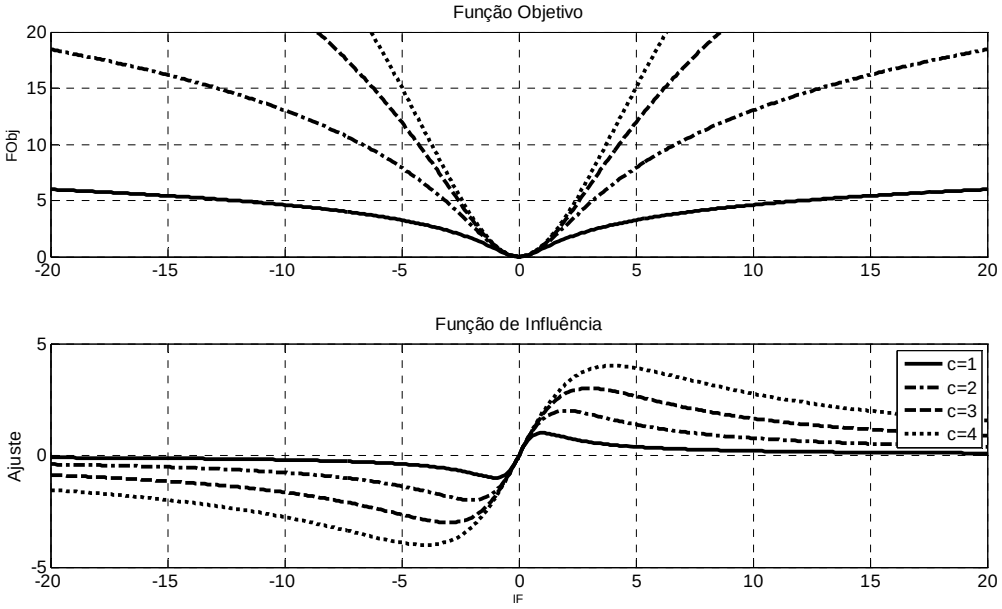

Figura 3.3: Influência do parâmetro *c* na função de Cauchy e na IF.

3.2.4 Função Fair

A função *Fair* foi construída como uma mistura entre os estimadores L_p – um caso especial de M-*estimadores* robusto (Butler et al., 1990). Ao invés de se estimar o somatório do erro quadrático, estimam-se as medidas minimizando a seguinte função objetivo:

$$\sum_{i=1}^{m} \left| y_i - f(x_i, \theta) \right|^p \tag{3.47}$$

onde o *mínimo desvio absoluto* corresponde à $p = 1$ e o *mínimo quadrado ordinário* corresponde à $p = 2$. Valores de $p<2$ estão associados com caudas mais finas que a distribuição normal. A função *fair* e sua função de influência são dadas por (Özyurt e Pike, 2004):

$$J_F = c^2 \left[\frac{|e_i|}{c\sigma_i} - ln\left(1 + \frac{|e_i|}{c\sigma_i} \right) \right] \tag{3.48}$$

$$IF_F = \frac{\partial J_F}{\partial \varepsilon} = sign(\varepsilon_i) \frac{1}{\sigma_i} \left(c - \frac{c^2}{\frac{|\varepsilon_i|}{\sigma_i} + c} \right) \tag{3.49}$$

E estas são mostradas na Figura 3.8, comparadas com a função objetivo normal.

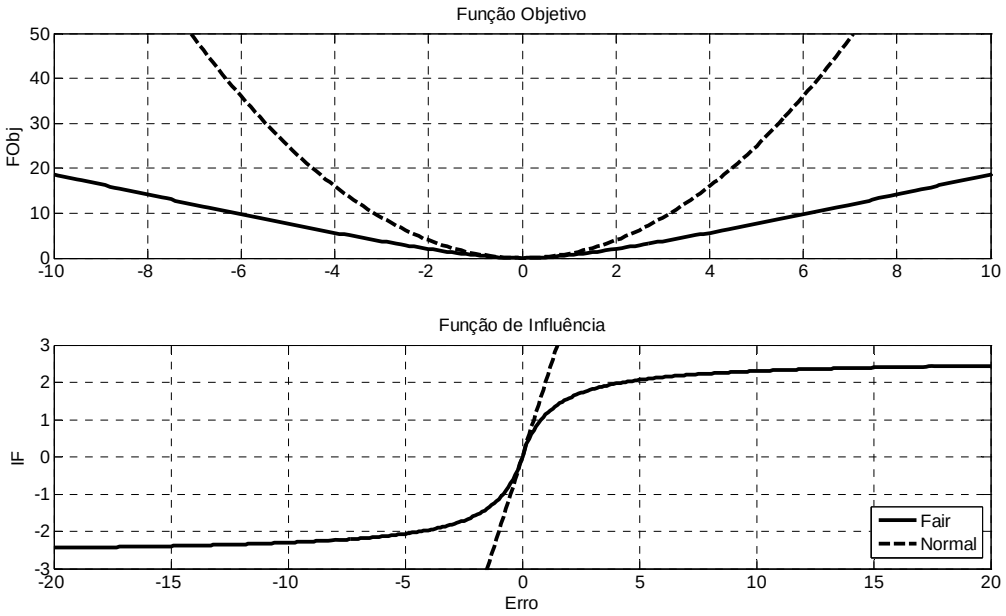

Figura 3.4: Comparação a Distribuição Normal e a Função *fair* (c=1,3).

Nota-se que a função de influência é limitada pelo parâmetro de sintonia *c* de maneira que $-c < IF_F(\varepsilon) < c$. Isto gera uma penalidade linear para erros grandes, mostrada na Figura 3.9. Pode-se notar que à medida que o parâmetro *c* aumenta, a função se aproxima da distribuição normal e o ajuste aumenta mais rapidamente em função do erro.

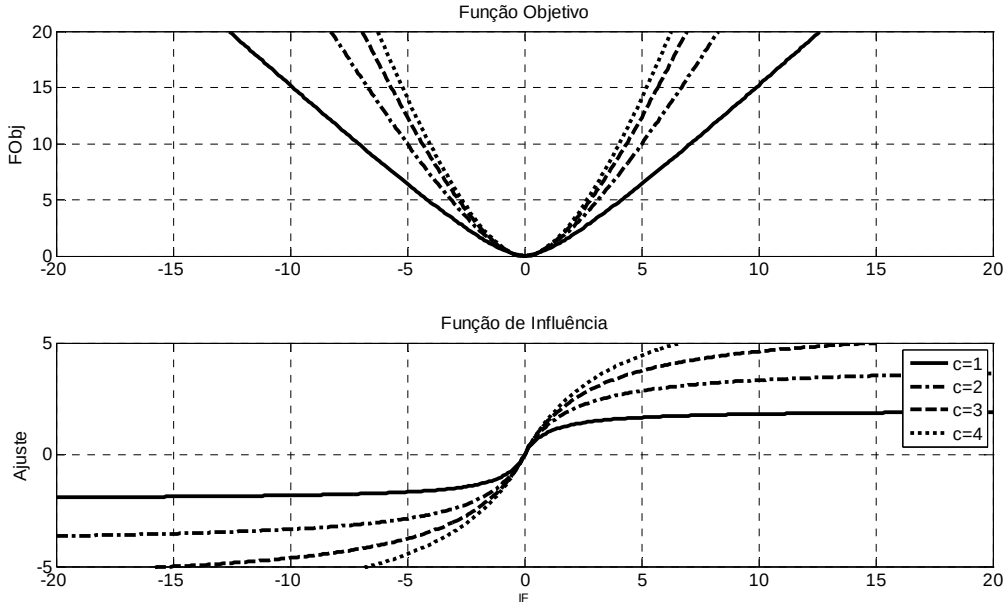

Figura 3.5: Influência do parâmetro *c* na distribuição Fair.

3.2.5 Solução do Problema de Reconciliação Robusta usando Programação Quadrática

Para a solução dos problemas de reconciliação robusta, visto que não é possível a obtenção da solução analítica, utilizou-se a implementação da função *fmincon* do *Matlab* para obtenção da solução através de técnica de programação quadrática seqüencial. Segundo Tjoa e Bieler (1991), este algoritmo apresenta bom desempenho para solução de problemas de reconciliação e é sugerido e utilizado pelos autores.

A programação quadrática seqüencial busca resolver as equações de KKT, ou condições necessárias de primeira ordem, permitindo a solução de problemas de otimização para funções objetivo não lineares e envolve a solução de uma seqüência de problemas de programação quadrática. Assim, partindo do problema de otimização geral,

$$\begin{aligned} &\underset{x}{Min}\ \ S(x) \\ &sujeito\ \grave{a} \\ &h(x) = 0 \\ &g(x) \le 0 \\ &x^L \le x \le x^U \end{aligned} \tag{3.50}$$

O problema de otimização é aproximado por restrições linearizadas e função objetivo quadrática e resolvido sucessivamente para encontrar as direções de busca d^k a partir da programação quadrática:

$$\begin{aligned} &\underset{d}{Min}\ \frac{1}{2} d^T H\left(x^k, \mu^k, \lambda^k\right) d + \nabla^T S\left(x^k\right) d \\ &sujeito\ a \\ &h\left(x^k\right) + \nabla^T h\left(x^k\right) d = 0 \\ &g\left(x^k\right) + \nabla^T g\left(x^k\right) d \le 0 \end{aligned} \tag{3.51}$$

Para gerar o problema de programação quadrática são feitas aproximações da Hessiana da função de Lagrange usando métodos de atualização tipo *quasi-Newton*. A solução de cada subproblema produz uma direção de busca d^k e após é realizada uma busca nesta direção (busca em linha) para obter o tamanho do passo ótimo de modo a obter o próximo ponto x_{k+1} a partir do ponto atual x_k.

3.3 Métodos para redução de tamanho do problema de RD

Apresentados os métodos para solução do problema de reconciliação utilizados neste trabalho, é interessante, neste momento, mostrar as maneiras de lidar com os casos especiais que podem aparecer quando as técnicas são aplicadas. Muitos processos industriais são complexamente integrados do ponto de vista mássico e energético. Por razões de custo, conveniência ou até razões técnicas, nem todas as variáveis são medidas. Assim, estas podem ser estimadas a partir das outras equações via cálculo de balanços.

A estimação das variáveis não medidas depende da estrutura do processo (topologicamente falando) e da localização dos sensores existentes. Tipicamente, existe um conjunto incompleto de instrumentos e as variáveis não medidas podem ser divididas em dois grupos: variáveis observáveis e não observáveis (aquelas que não podem ser obtidas a partir das variáveis medidas existentes). As variáveis medidas do processo podem ser divididas em redundantes e não redundantes (ou seja, aquelas variáveis que deixam de ser observáveis caso a sua observação seja suprimida).

A RD usa a redundância nas medidas para obtenção das estimativas e desta forma deve contemplar somente as variáveis redundantes. As variáveis não redundantes não adicionam informação ao problema e não serão ajustadas durante o procedimento. Desta forma, é necessário que o problema seja dividido e resolvido por partes. A primeira parte refere-se ao problema de reconciliação propriamente dito, onde só devem participar as variáveis adequadas para tal fim. Na segunda parte, as variáveis não-medidas podem ser estimadas (como já mencionado anteriormente, este é conhecido como problema de cooptação) e para isto, só devem ser levadas em conta as variáveis que são observáveis. Sobre as variáveis que não participaram de nenhum dos dois problemas nada se pode afirmar. As variáveis medidas não redundantes são somente determinadas e as variáveis não medidas não observáveis não podem ser determinadas, a menos que sejam inseridas outras medições.

Nesta dissertação este método é utilizado no contexto do tratamento dos erros grosseiros, onde algumas técnicas envolvem a eliminação da medição suspeita de conter erro grosseiro e aplicação de algum teste estatístico no conjunto de medidas resultante. Se o conjunto de observações resultante não passar no teste estatístico este é um bom indicativo da presença de erro grosseiro na variável suspeita e eliminada. O método mais utilizado para a separação do problema de RD é o *método da matriz de projeção* e será apresentado a seguir.

3.3.1 Método da Matriz de Projeção

O método da matriz de projeção de Crowe (1983) foi proposto para reduzir o problema de RD, obtendo-se a solução analítica usando projeções para separar as variáveis. A projeção é usada para separar o conjunto das variáveis medidas das não medidas, resolver o problema de RD para as variáveis disponíveis e projetar a solução para encontrar as variáveis não medidas, se for possível. Quando existem variáveis não medidas, ou se quer descartar alguma medida para testar a existência de erros grosseiros, pode-se separar o problema da seguinte forma:

$$A_x y_x + A_u y_u = 0 \tag{3.52}$$

Onde A_x e y_x são os coeficientes das equações de balanço das variáveis medidas e o conjunto das variáveis medidas e A_u e y_u são referentes às variáveis não medidas. Caso as variáveis não medidas sejam observáveis, o conjunto y_u poderá ser determinado posteriormente. Fazendo com que a matriz P seja:

$$PA_u = 0 \tag{3.53}$$

A matriz P é ortogonal ao espaço coluna de A_u. Multiplicando as equações de restrições (Equação 3.52):

$$PA_x y_x = 0 \tag{3.54}$$

Se P é uma matriz de projeção, então a matriz reduzida das restrições A_r (onde só existem as variáveis medidas) é definida como sendo:

$$A_r = PA_x \tag{3.55}$$

As estimativas das variáveis medidas reconciliadas são obtidas substituindo-se a relação dada pela Equação 3.55 na Equação 3.23. Esta substituição resulta em:

$$\hat{x}_x = y_x - W^{-1}PA_x^T(PA_xW^{-1}PA_x^T)^{-1}(PA_xy_x) \tag{3.56}$$

E se as variáveis não medidas (y_u) forem observáveis, as suas estimativas são obtidas por:

$$\hat{x}_u = -[\,A_u{}^T (A_x W^{-1} A_u^T)^{-1} A_u{}^T\,]^{-1} A_u{}^T (A_x W^{-1} A_x^T)^{-1} A_x \hat{x}_x \tag{3.57}$$

Existem na literatura diferentes métodos de construção da matriz de projeção. O método mais utilizado é o da *Decomposição QR* da matriz A_u. Este será apresentada na seção que segue.

3.3.2 Obtenção da matriz de Projeção: Decomposição QR

Para a construção da matriz de projeção, pode-se utilizar a decomposição QR da matriz de variáveis não medidas $\mathrm{A_u}$. Esta consiste em, dada uma matriz M (m x n) com $m \geq n$ e n colunas linearmente independentes, então existe uma matriz Q única (m x m) onde,

$$\begin{aligned} &Q^T Q = D \\ &D = diag(d_1,...,d_k);\;\; d_k > 0, \qquad k = 1,...,n \end{aligned} \tag{3.58}$$

E existe uma matriz triangular superior única (m x n) R, com $\mathrm{R_{kk}} = 1, \mathrm{k} = 1,..., \mathrm{n}$ de maneira que:

$$M = QR \tag{3.59}$$

O objetivo de qualquer método de ortogonalização é encontrar uma base ortogonal para $R(M)$. De fato, se $R(M) = R(Q_1)$, onde $Q_1 = [q_1,...,q_r]$ tem colunas ortonormais, então $M = Q_1 S$ para alguns $S \in \Re^{r\,x\,n}$. Mas se a matriz M tem posto deficiente então pelo menos uma linha de R é nula e a decomposição QR de uma matriz M pode não produzir uma base ortogonal para $R(M)$. Neste caso, a decomposição QR pode ser modificada,

$$M\Pi = [Q_1 \quad Q_2] \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \tag{3.60}$$

Onde r = posto(M), Q é ortogonal, R_{11} é triangular superior e $\prod$ é a matriz de permutação. Se,

$$M\Pi = [a_{c1},...,a_{ck}] \tag{3.61}$$

$$Q = [q_1,...,q_m] \tag{3.62}$$

Então para k = 1,..., n

$$a_{ck} = \sum_{i=1}^{min\{r,k\}} r_{ik} q_i \in span\{q_1,...,q_r\} \tag{3.63}$$

E que para qualquer vetor que satisfaça $Mx = b$,

$$\Pi^T x = \begin{bmatrix} s \\ z \end{bmatrix} \quad e \quad Q^T b = \begin{bmatrix} i \\ 1 \end{bmatrix} \tag{3.64}$$

Onde *s* e *i* são vetores de dimensão *r*, *z* é um vetor de dimensão (*n-r)* e 1 é um vetor de dimensão (*m-r*). Agora, voltando à decomposição da matriz A_u e considerando que as colunas da matriz A_u são linearmente independentes, é possível decompô-la como:

$$A_u = QR\Pi_U = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \Pi_u \tag{3.65}$$

Onde Π_u é a matriz de permutação (as colunas de Π_u são as colunas permutadas da matriz identidade). R_1 é uma matriz não singular, triangular superior *p* x *p*. *Q* é a matriz ortonormal (*m* x *m)*, de maneira que, $Q^T Q = I$. As colunas de *Q* formam a base para o espaço m-dimensional, enquanto a matriz R_1 representa as *p* colunas de A_u em termos dos primeiros *p* vetores bases, Q_1. Por *Q* ser ortogonal, a matriz Q_2 tem a propriedade

$$Q_2^T A_u = Q_2^T \begin{bmatrix} Q_1 Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \Pi_u = \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \Pi_u = 0 \tag{3.66}$$

Na equação 3.66 fica claro que a matriz Q_2^T é a matriz de projeção *P* desejada. Para a obtenção das variáveis não medidas, usando a fatoração QR, podemos escrever:

$$A_x y_x + QR\Pi_u y_u = 0 \tag{3.67}$$

Onde Π_u é o vetor y_u reordenado. Multiplicando a equação por Q^T temos:

$$Q^T A_x y_x + R\Pi_u y_u = 0 \tag{3.68}$$

ou

$$-Q^T A_x y_x = R\Pi_u y_u \tag{3.69}$$

Usando a definição de A_u

$$-\begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} A_x y_x = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \Pi_u y_u \tag{3.70}$$

Ou

$$-Q_1^T A_x y_x = R_1 \Pi_u y_u \tag{3.71}$$

Substituindo Q_2^T na Equação 3.56 para obter as estimativas das variáveis reconciliadas temos:

$$\hat{x}_x = y_x - W(Q_2^T A_x)^T [(Q_2^T A_x)W(Q_2^T A_x)^T]^{-1}(Q_2^T A_x)y_x \tag{3.72}$$

Como R_1 é uma matriz *p x p* triangular superior podemos resolver a equação 3.71 para obter as estimativas de y_u. A solução pode ser expressa como:

$$\Pi_u x_u = -R_1^{-1} Q_1^T A_x y_x \tag{3.73}$$

Ou ainda pela equação equivalente abaixo:

$$\hat{x}_u = -\left[A_u^T (A_x W^{-1} A_x^T) A_u^T\right]^{-1} A_u^T (A_x W^{-1} A_x^T)^{-1} A_x \hat{x}_x \tag{3.74}$$

Desta forma é possível resolver o problema de reconciliação, mesmo que na presença de variáveis não medidas, contanto que estas sejam observáveis.

3.4 Classificação de variáveis

Antes de realizar a reconciliação de dados, é interessante que sejam identificadas quais as variáveis são não observáveis ou não redundantes. No contexto do tratamento de erros grosseiros, se estes estiverem presentes em medidas não redundantes, podem levar a valores não observáveis e a solução obtida pode não ser única e as suas estimativas sem sentido (Arora e Biegler, 2001).

Aproveitando a demonstração do método para decomposição do problema de RD, será introduzido um método para a classificação das variáveis, baseado na mesma decomposição QR. Existem outros métodos utilizados para realizar esta tarefa, já citados na revisão bibliográfica e mas, por sua afinidade com o método de obtenção da matriz de projeção, este foi escolhido para ser utilizado nesta dissertação.

3.4.1 Classificação das variáveis via Decomposição QR

A decomposição QR, usada para obtenção da matriz de projeção, também pode ser usada para obter informações sobre as condições de estimabilidade das variáveis, permitindo uma classificação direta (Romagnoli e Sanchez, 2000). De acordo com o que já foi apresentado, a decomposição QR da matriz A_u permite que sejam encontradas matrizes *Q* e *R* e a matriz de permutação $\prod_u$ de maneira que:

$$A_u \Pi_u = QR \tag{3.75}$$

As matrizes *Q* e *R* podem ser divididas em:

$$Q = [Q_1 \quad Q_2] \tag{3.76}$$

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \tag{3.77}$$

Onde r = posto (A_u) = posto(R_{11}). A matriz Q é ortonormal e R_{11} é uma matriz triangular superior não-singular de dimensão r. Da mesma maneira, as variáveis não-medidas podem ser separadas em dois subconjuntos,

$$\Pi_u^T x_u = \begin{bmatrix} x_{ur} \\ x_{u(n-r)} \end{bmatrix} \tag{3.78}$$

Se as restrições forem multiplicadas por $Q^T = Q^{-1}$ tem-se:

$$\begin{bmatrix} Q_1^T A_1 & R_{u1} & R_{u2} \\ Q_2^T A_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ x_{ur} \\ x_{u(n-r)} \end{bmatrix} = 0 \tag{3.79}$$

Então as r primeiras equações de x_{ur} podem ser escritas em termos das outras variáveis,

$$x_{ur} = -R_{11}^{-1} Q_{11}^T A_x x_x - R_{11}^{-1} R_{12} x_{u(n-r)} \tag{3.80}$$

A matriz R, obtida da decomposição de A_u, contém informações topológicas do sistema em termos das variáveis disponíveis e a matriz de permutação permite a classificação das variáveis não medidas do processo. As variáveis do subconjunto $x_{u(n-r)}$ correspondem ao número mínimo e localização das medidas necessárias para que o sistema satisfaça as condições de estimabilidade (que todas as variáveis não medidas possam ser determinadas). Assim tem-se regras para realizar a classificação das variáveis de processo, propostas por Sanchéz e Romagnoli (1996):

1) Se o posto de $R = r = n$, onde n é o número de variáveis não medidas, então todas as variáveis não medidas podem ser estimadas a partir das informações disponíveis;

2) Se o posto de $R = r < n$, então pelo menos $(n\text{-}r)$ variáveis não podem ser calculadas a partir das informações disponíveis;

3) Da matriz Π_u pode-se obter a classificação das variáveis não medidas de processo. As variáveis do subconjunto $x_{u(n-r)}$ correspondem ao número mínimo e a localização das medidas necessárias para que o sistema satisfaça a condição de estimabilidade – que todas as variáveis não medidas possam ser determinadas;

4) As colunas da matriz Q_2^T geram o espaço nulo de A_u, ou seja,

$$Q_2^T A_u = 0 \tag{3.81}$$

Nada se pode dizer sobre o subconjunto x_{ur}, pois estas variáveis podem ser calculadas diretamente das medidas disponíveis e algumas dependem da escolha das variáveis de $x_{u(n-r)}$. Mas é possível classificar estas variáveis a partir da decomposição QR. Nota-se que, na equação 3.80, se o lado direito for igual à zero, então todas as x_{ur} variáveis correspondentes podem ser calculadas das informações disponíveis. Para classificar as x_{ur} variáveis restantes, deve-se olhar para as linhas da seguinte matriz:

$$R_{IU} = R_{11}^{-1} R_{12} \tag{3.82}$$

Consequentemente,

5) As variáveis do subconjunto x_{ur} são ditas estimáveis se a linha correspondente na matriz R_{IU} for igual à zero. E o subconjunto x_{ur} é dito não estimável, caso contrário;

6) A classificação das variáveis medidas pode ser feita examinando a matriz associada como as equações da reconciliação. As colunas de zero de A ou A_x correspondem às variáveis que não participam na reconciliação e então estas são não redundantes. As colunas restantes correspondem às medidas redundantes.

Da mesma forma que o método para separação do problema de RD, a classificação das variáveis é utilizada no contexto da detecção de erros grosseiros. Esta é utilizada em conjunto com os métodos de detecção, e muitas vezes como um critério de parada forçada do algoritmo e a sua utilização específica será mostrada ao longo da demonstração teórica dos mesmos. A apresentação desta, neste momento, é feita para facilitar o entendimento teórico e aproveitar a seqüência de conceitos apresentados.

Capítulo 4

Detecção de Erros Grosseiros

Um dos tópicos mais explorados na área de reconciliação de dados é a Detecção de Erros Grosseiros (DEG). É interessante não somente ter a localização exata do erro para ajudar na manutenção dos instrumentos e monitorar vazamentos, como também ter uma estimativa de sua magnitude, para o fechamento dos balanços de massa (Jiang e Bagajewicz, 1999). Segundo Romagnoli e Sanchez (2000) existem diferentes maneiras de se identificar erros grosseiros, entre eles,

- Com análise teórica de todas as fontes de erros grosseiros;
- Com medições de variáveis de processo obtidas por dois métodos com precisões diferentes;
- Conferindo se as equações de balanço são satisfeitas.

Este último é particularmente atrativo, pois é relativamente simples e é baseado em relações de validade absoluta – as leis de conservação de massa e energia (Romagnoli e Sanchez, 2000). A detecção de erros grosseiros é parte fundamental de um sistema de reconciliação de dados. Este, geralmente, baseia-se na realização de testes de hipóteses sobre o conjunto de dados ou sobre transformações lineares destes. Aproveitam-se as propriedades estatísticas da modelagem dos erros aleatórios para correta detecção, e infere-se sobre o conjunto de dados, usando a redundância gerada pela utilização dos modelos e alguma relação conhecida dos dados como, por exemplo:

$$t = \frac{\hat{E}\left(y_i - x_i\right)}{\hat{\sigma}} \tag{4.1}$$

O teste de hipóteses pode ser,

Hipótese Nula: $H_0 : \hat{E}\left(y_i - x_i\right) = 0$ (4.2)

Hipótese Alternativa: $H_1 : \hat{E}(y_i - x_i) \neq 0$ (4.3)

Sob H_0, o teste estatístico é comparado com um valor de referência (*valor crítico*) que é um *ponto de corte* da distribuição definida sob certo nível de confiança. Assim, a hipótese nula é aceita ou rejeitada. Se uma função densidade de probabilidade pode ser assumida para *t*, grandes valores de *t* descreverão situações menos esperadas e darão a prova para que H_1 seja verdadeira, i.e., a existência de erros grosseiros. Desta forma, o teste de hipótese é o processo de decisão.

Do ponto de vista teórico, pode-se pensar no processo de decisão como uma divisão do espaço $\Re^n$ em duas regiões – R_0 e R_1 – conhecidas como *região de aceitação* e *região de rejeição* (*ou crítica*), respectivamente. Sendo x=[x_1, ... ,x_n] o vetor observado, então se $x \in R_0$, decide-se por H_0, caso contrário, decide-se por H_1, existindo 4 possibilidades de decisão:

1. Aceitar H_0 quando H_0 é verdadeira: *Decisão Correta.*

2. Rejeitar H_0 quando H_0 é verdadeira: *Decisão Errada*

3. Aceitar H_1 quando H_1 é verdadeira: *Decisão Correta*

4. Rejeitar H_1 quando H_1 é verdadeira: *Decisão Errada*

Destas, existem duas possibilidades relacionadas a decisões equivocadas. A rejeição da hipótese nula quando esta é verdadeira é chamada de Erro Tipo I (ou ainda *Falso Alarme*). A rejeição da hipótese alternativa quando esta é verdadeira é chamada de Erro Tipo II (*Não Detecção*). É interessante tratar os erros tipo I e II em função de suas probabilidades de realização. Assim, designando D_0 como a decisão de não aceitar H_1 e D_1 como a decisão de aceitar H_1. Define-se:

Probabilidade de Erro Tipo I: $P_I = p(D_1|H_0) = p\{x \in R_1; H_0\}$ (4.4)

Probabilidade de Erro Tipo II: $P_{II} = p(D_0|H_1) = p\{x \in R_0; H_1\}$ (4.5)

A probabilidade de erro tipo I também é conhecida por α e este representa o nível de crença ou confiança que se quer ter quando o teste é definido. O nível de confiança define a probabilidade de se cometer erros tipo I (ou pelo menos é o limite superior). Já a probabilidade de erro tipo II é também chamada de β e define-se como o *Poder do teste estatístico* da probabilidade complementar à β.

$$Poder\ do\ teste = (1 - \beta) \quad (4.6)$$

Pode-se notar que α e β representam as probabilidades dos eventos a partir do mesmo problema de decisão e logo não são independentes uma da outra. Geralmente a diminuição em

um tipo de erro leva ao aumento do outro. Na Figura 4.1, é mostrada a relação entre o erro tipo I e o erro tipo II para um teste univariável, onde se observa que ambos podem ser reduzidos a partir da escolha apropriada de um valor crítico $Z_{1-\alpha/2}$ e à medida que o tamanho do erro grosseiro (δ) aumenta. Desta forma, o desempenho dos métodos de detecção é fortemente afetado pela relação entre o desvio padrão do erro aleatório presente na medição e o tamanho do erro grosseiro (Chen et al., 2001).

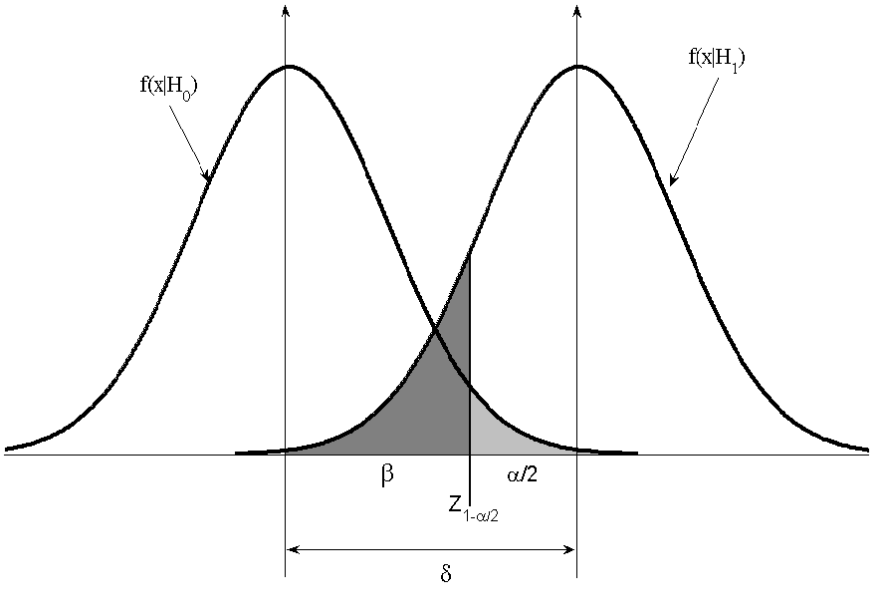

Figura 4.1: Relação entre erro tipo I e erro tipo II.

Quando o teste é realizado simultaneamente para mais de uma variável, o nível de significância é modificado. Quando múltiplos testes são realizados, a probabilidade de se cometer erro tipo I é maior do que o valor especificado por α, e então uma modificação é necessária para que a região de rejeição seja reduzida e esta probabilidade passe a ser controlada novamente. Sendo α o nível de significância global, e *m* o número de variáveis independentes, então o nível de significância para cada variável é:

$$\alpha^* = 1 - (1 - \alpha)^{1/m} \tag{4.7}$$

O valor de referência é dado por (1- α*/2), isto é, o ponto de corte da distribuição normal padrão. Isto garante que a probabilidade de que qualquer um dos testes realizados simultaneamente seja rejeitado sob H_0 seja menor ou igual à α. Para grandes valores de *m* ou quando existe correlação entre as variáveis, Rollings e Davis (1992) propuseram o uso do valor crítico baseado no intervalo de confiança de Bonferroni, que é dado por:

$$\alpha^* = \alpha / m \tag{4.8}$$

Sidak (1967) provou que o valor crítico t_c dá o limite superior exato, ou seja:

$$P[t \geq t_c] \equiv \frac{\alpha}{m} \tag{4.9}$$

Conforme já citado no Capítulo 3, uma estratégia de detecção deve ter as habilidades de: detectar, localizar e estimar o erro grosseiro. É importante frisar que nem todos os métodos apresentam todas estas habilidades e nem conseguem lidar com todos os tipos de erros grosseiros, pois esta depende do tipo de teste de hipótese realizado e do modelo utilizado.

Na detecção, o modelo é utilizado no sentido de que se as variáveis seguem certa distribuição, então os valores medidos e suas transformações lineares (modelo linear) também a seguem. Os testes mais populares são baseados em duas transformações lineares: o resíduo das restrições (modelo do processo) ou sobre o ajuste realizado em uma etapa de reconciliação prévia. Portanto, utilizando as restrições tem-se,

$$r = Ay = A(x + \varepsilon) \tag{4.10}$$

Como r é uma transformação linear das variáveis medidas, este segue, assim como o erro presente na medição, uma distribuição normal multivariada, N(0,V), e a sua variância segue uma distribuição qui-quadrado com υ graus de liberdade (onde υ é posto da matriz A). Não existindo erros grosseiros, a esperança para r e sua variância V são:

$$E[r] = E[Ay] = AE[\varepsilon] = 0 \tag{4.11}$$

$$V = cov[r] = E[(A\varepsilon)(A\varepsilon)^T] = A\,E(\varepsilon\varepsilon^T)A^T = AWA^T \tag{4.12}$$

Desta maneira, pode-se realizar o teste de hipótese da Equação 4.1, adaptado para as restrições. Este é um dos testes mais populares na literatura de RD e DEG e conhecido como *teste da restrição (ou nodal)*. Da mesma forma pode ser estabelecido um teste baseado em uma transformação diferente das restrições, seguindo a mesma filosofia, baseado em PCA e chamado de *teste da componente principal*.

A segunda maneira mais popular de se aplicar o teste de hipótese é baseada nos ajustes realizados em uma etapa de reconciliação de dados prévia. O ajuste obtido na RD é a estimativa do erro aleatório e definido como:

$$a = y - \hat{x} = \hat{\varepsilon} \tag{4.13}$$

$$E[\hat{\varepsilon}] = 0 \tag{4.14}$$

Usando a solução analítica do problema de reconciliação, os ajustes são dados por:

$$a = WA^T V^{-1} r \tag{4.15}$$

Ou ainda, substituindo a relação para V e r:

$$a = WA^T (AWA^T)^{-1} Ay \tag{4.16}$$

E a sua variância é dada por:

$$W_a = cov(a) = WA^T V^{-1} W \tag{4.17}$$

Se $\varepsilon \sim N(0,W)$, então $a \sim N(0,W_a)$ visto que estes são somente uma transformação linear dos erros presentes nas medidas e a variância W_a segue uma distribuição qui-quadrado com υ graus de liberdade (υ = posto(A)). Portanto, pode-se realizar o teste de hipótese da Equação 4.1, adaptado para as medições. Este é conhecido como *teste da medição*.

4.1 Testes para detecção de um único erro grosseiro

Nesta seção serão apresentados os cinco testes mais populares para detecção de erros grosseiros. Estes são utilizados nas estratégias de detecção de múltiplos erros, sendo que podem ser apresentadas diferentes variações. Alguns destes testes são equivalentes, mas apresentam filosofias ou princípios diferentes. Estes são:

- ✓ Teste Global: tem por objetivo avaliar todo o conjunto de dados. Baseia-se no valor da função objetivo do problema de otimização e é o teste mais utilizado na literatura, visto sua simplicidade.

- ✓ Teste da Restrição (ou Nodal): avalia cada uma das restrições em relação à função objetivo do problema de reconciliação. É simples, mas indica somente a restrição que contém o erro grosseiro. Além deste, é apresentada a variação de máxima potência, que pode ser utilizada quando a matriz de variância-covariância não é diagonal.

- ✓ Teste da Medição: utiliza como referência uma etapa de reconciliação prévia para testar cada uma das variáveis medidas. É o teste mais utilizado na literatura, em conjunto com o teste global. Os ajustes realizados nas variáveis medidas são comparados com aquele que deveria ser feito caso só existissem erros aleatórios nas medidas. Além deste, é apresentada uma variação de máxima potência.

- ✓ Teste GLR: baseia-se no princípio da máxima verossimilhança para determinação das variáveis com erros grosseiros. Este método apresenta o diferencial de poder detectar diferentes tipos de erros grosseiros, dependendo de como é escolhido o modelo da medição ou da restrição. Quando se considera somente a presença de erros do tipo bias, ele é equivalente ao teste da medição.

- ✓ Teste PCA: é uma formulação alternativa, que pode ser equivalente tanto ao teste da medição quanto ao teste da restrição e pode lidar com variáveis correlacionadas.

Os testes, separadamente, nem sempre podem detectar, identificar e estimar os erros grosseiros. Por exemplo, o teste global, só realiza a etapa de detecção e nada mais pode ser concluído sobre o conjunto de dados. O teste nodal dá uma ideia da localização indicando qual a restrição que contém o erro, mas não identifica exatamente a variável. Já os testes da medição, GLR e PCA indicam exatamente a variável que contém o erro, O GLR ainda pode lidar com erros grosseiros do tipo vazamento, dependendo do modelo utilizado.

4.1.1 Teste Global (GT)

O GT ("Global Test"), proposto por Ripps (1965), é baseado no vetor de resíduos dos balanços *r*, representado pela Equação 4.10. Quando na presença de erros grosseiros, o vetor *r* reflete o grau de violação das restrições de processo. O teste estatístico usado é o seguinte:

$$\gamma = r^T V^{-1} r = (Ay)^T V^{-1} (Ay) \tag{4.18}$$

Se a matriz A tem posto igual ao número de restrições, sob H_0, γ segue uma distribuição qui-quadrado com υ graus de liberdades, onde υ = posto(A). O valor crítico da distribuição qui-quadrado é obtido de maneira que,

$$P\{\gamma \geq \chi^2_{1-\alpha,\upsilon}\} = \alpha \tag{4.19}$$

Onde α é o grau de confiança. Se $\gamma \geq \chi^2_{1-\alpha,\upsilon}$, então H_0 é rejeitada e o erro grosseiro é detectado. O fluxograma do algoritmo é apresentado a seguir:

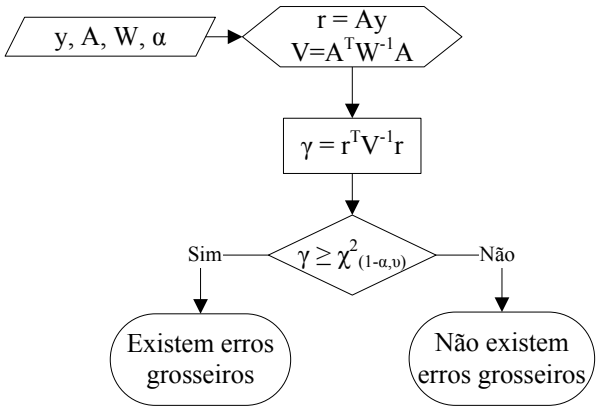

Figura 4.2: Fluxograma do Teste Global.

O Teste Global reflete o valor da função objetivo de mínimos quadrados ponderados. Este teste não dá nenhuma informação sobre a causa ou localização do erro grosseiro e, por isto, existe a necessidade de aplicar outros testes ao conjunto de dados. A principal vantagem

é a extrema simplicidade e por este motivo sempre está presente em praticamente todas as estratégias de detecção.

4.1.2 Teste da Medição (MT)

O MT ("Measurement Test), proposto por Mah e Tamhane (1982), investiga os ajustes da medição *a* gerados pelo problema de reconciliação. De modo a criar uma estimativa de referência, uma reconciliação prévia é realizada e estes ajustes são testados. Assim o teste é realizado em função da estatística Z_a, definida como:

$$Z_{a,i} = \frac{|a_i|}{\sqrt{(W_a)_{ii}}} \qquad i = 1,2,...,m \tag{4.20}$$

Onde *a* é o vetor de ajustes obtido pela Equação 4.16 e $(W_a)_{ii}$ é o elemento da diagonal da matriz de variância-covariância dos ajustes, obtida pela Equação 4.17. Sob H_0, Z_a segue uma distribuição normal padrão, N(0, 1). Então,

$$\text{Hipótese nula:} \qquad H_0 : Za_i = 0 \tag{4.21}$$

$$\text{Hipótese alternativa:} \quad H_1 : Za_i \neq 0 \tag{4.22}$$

Como são realizados múltiplos testes simultaneamente, o nível de confiança α é modificado de acordo com a Equação 4.7 e o tem-se o valor crítico:

$$Za_c = Z_{(1-\alpha^*/2)} \tag{4.23}$$

Assim, os valores de $Za_i > Za_c$ farão com que a hipótese nula seja rejeitada e estas variáveis serão detectadas como contendo erros grosseiros. Na Figura 4.3 é mostrado o fluxograma do algoritmo.Este teste é interessante por sua simplicidade e facilidade de cálculo. A maior desvantagem é que quando existem múltiplos erros grosseiros, a etapa de reconciliação espalhará estes erros por todo o conjunto de dados, piorando a precisão das variáveis que não apresentavam erros. Assim, quando o teste é aplicado, muitas variáveis que não contém erro serão detectadas, implicando no cometimento de erros tipo I.

Mah e Tamhane (1982) demonstraram que, quando a matriz de covariância é não diagonal, existe outra maneira de calcular o vetor ajustes, conferindo a este a propriedade de máxima potência para detecção. Este é conhecido com *Teste da Medição de Máxima Potência (MTMP)*. O vetor *d* é obtido pela multiplicação da matriz de covariância pelo vetor dos ajustes *a* como segue:

$$d = W^{-1}a \tag{4.24}$$

Sob a hipótese nula H_0, *d* também segue a distribuição normal com média zero e covariância conhecida W_d e pode-se definir a matriz de covariância de *d* como sendo:

$$W_d = cov[d] = A^T (AWA^T)^{-1} A \tag{4.25}$$

Então foi proposto por Mah e Tamhane (1982) o seguinte teste estatístico:

$$Z_{d,i} = \frac{|d_i|}{\sqrt{(W_d)_{ii}}} \quad i = 1,2,...m \tag{4.26}$$

Utilizando o intervalo de confiança modificado, o valor crítico é dado por $Z_{d,c}$:

$$Z_{d,c} = Z_{(1-\alpha^*/2)} \tag{4.27}$$

O parâmetro α^* é escolhido conforme a Equação 4.7 garantindo assim que a probabilidade de *erro tipo I* seja menor ou igual a α.

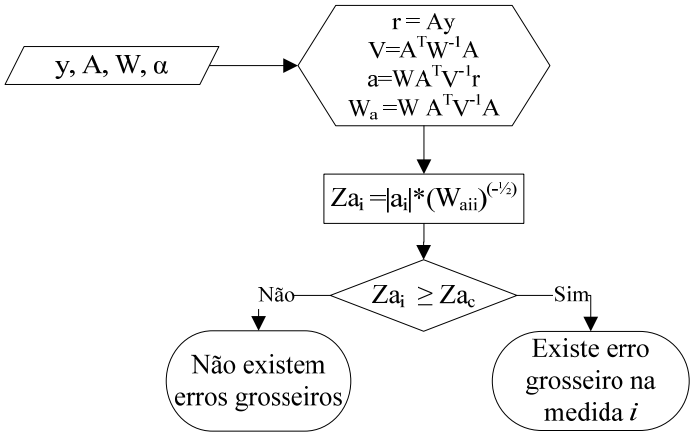

Figura 4.3: Fluxograma do Teste da Medição..

4.1.3 Teste da Restrição (NT)

O teste da restrição, ou nodal, foi proposto por Reilly (1963). Este testa os resíduos das equações de balanço frente a um critério estatístico, Zr_j, baseado no resíduo ao redor do nó j:

$$Zr_j = \frac{|r_j|}{\sqrt{V_{jj}}} \quad j = 1,2,...n \tag{4.28}$$

Onde Zr_j segue uma distribuição normal. O nível de significância α^* modificado, proposto por Mah a Tamhane (1982) pode ser usado e é dado por:

$$\alpha^* = 1 - (1-\alpha)^{1/n} \tag{4.29}$$

O valor crítico é dado por Zr_c:

$$Zr_c = Z_{(1-\alpha*/2)} \tag{4.30}$$

Se $Zr_j > Zr_c$, então H_0 é rejeitada e o erro grosseiro é detectado na restrição j. Na Figura 4.4 é apresentado o fluxograma do algoritmo. Este é um teste muito popular na literatura, mas apresenta duas desvantagens. A primeira é que ele não define a exata localização do erro grosseiro e é necessário o uso de um teste combinado a este. A segunda desvantagem é o cancelamento dos erros grosseiros pertencentes à mesma restrição. Isto será demonstrado mais adiante, quando serão apresentadas as estratégias para detecção de múltiplos erros grosseiros.

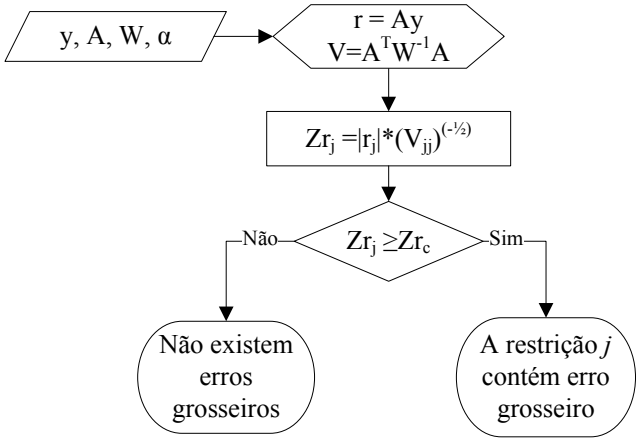

Figura 4.4: Fluxograma do algoritmo MT.

É possível obter outras formas de testes para as restrições usando transformação linear dos resíduos das restrições. Entretanto nem todas estas transformações possuem o mesmo poder de detecção. Crowe (1989) sugere uma forma particular do teste da restrição que tem a máxima potência de detecção. Este é conhecido como o *Teste da Restrição de Máxima Potência* (NTMP), cujo teste é calculado de maneira que:

$$Z^*_{r,j} = \frac{\left|[V^{-1}r]_j\right|}{\sqrt{[V^{-1}]_{jj}}} \quad j = 1,2,...n \tag{4.31}$$

O critério para comparação pode ser escolhido da mesma forma que para o teste convencional. Se existem erros grosseiros no processo, o máximo valor esperado entre os testes estatísticos gerados pela equação acima é maior do que o valor esperado pelos testes sugerido na Equação 4.28. Isto implica em que se existe um erro grosseiro, o teste da máxima potência tem maior probabilidade de detecção (Crowe, 1989; Romagnoli et al., 2000; Narashiman et al., 2000).

4.1.4 Teste da Razão de Máxima Verossimilhança Generalizada (GLR)

O teste GLR ("*Generalized Likelihood Ratio"*), proposto por Narashimhan et al. (1987), é baseado na estimação de máxima verossimilhança. Este apresenta a capacidade de detectar qualquer tipo de erro grosseiro que possa ser modelado matematicamente. Uma das maiores vantagens deste método é que a identificação do erro grosseiro não pode ser confundida com saídas das condições de estado estacionário, o que, nos outros métodos, poderia ser identificado como erro grosseiro do tipo vazamento. Além disto, uma estimativa para o valor do erro grosseiro é obtida ao longo do processo de estimação, o que pode ser muito útil para a avaliação do impacto deste erro no sistema. O modelo para bias de magnitude *b* desconhecida no instrumento *i* é dado por:

$$y = x + \varepsilon + be_i \tag{4.32}$$

Onde e_i é um vetor com 1 na posição *i* e zeros nas demais. Se o interesse for detectar a presença de vazamentos, então o modelo do processo é baseado em uma modificação das restrições,

$$Ay - bm_j = 0 \tag{4.33}$$

Onde m_j é um vetor com 1 na posição relativa as restrições relacionadas com o nó *j* e zeros nas demais. Em Narashiman et al. (1987), estes aconselham que o teste para vazamento deva ser feito somente se existir uma grande possibilidade de tê-los. Além disto, os valores unitários são escolhidos como sendo positivos e quando estimados valores negativos deve-se encarar como condições não estacionárias de processo.

O algoritmo de detecção consiste em, dadas as medições, quer-se determinar se existem erros grosseiros, onde estão localizados e identificar a fonte ou causa. Primeiro se considera o caso em que no mínimo existe *um* erro grosseiro. Se não existem erros grosseiros então a Equação 4.10 é válida. Ao contrário, quando um erro grosseiro de bias ou vazamento está presente, pode-se dizer que:

$$E[r] = bf_k \tag{4.34}$$

$$f_k = \begin{cases} Ae_i & \text{para bias na medida } i \\ m_j & \text{para vazamento no nó } j \end{cases} \tag{4.35}$$

Definindo-se μ como o valor esperado e desconhecido de *r* (o vetor das restrições) podem-se formular as hipóteses para a detecção de erros grosseiros como:

$$H_0 : \mu = 0 \text{: Não existem erros grosseiros} \tag{4.36}$$

$$H_1 : \mu = bf_k \text{: Um erro grosseiro está presente em k} \tag{4.37}$$

O parâmetro b pode ser qualquer número real e o parâmetro f_k pode ser qualquer vetor do conjunto F dado por:

$$F = \{Ae_i, m_j, i = 1...n, j = 1...m\} \tag{4.38}$$

De maneira a testar as hipóteses dadas pelas equações anteriores e estimar os parâmetros desconhecidos se H_1 for aceita, faz-se uso do teste da máxima verossimilhança. Este é dado pela Equação 4.39. O supremo é obtido sobre todos os valores possíveis dos parâmetros presentes nas hipóteses.

$$\lambda = sup \frac{p\{r|H_1\}}{p\{r|H_0\}} \tag{4.39}$$

Usando a função densidade de probabilidade normal para r, a Equação 4.39 torna-se:

$$\lambda = sup \frac{exp\left[-\frac{1}{2}(r - bf_k)^T V^{-1}(r - bf_k)\right]}{exp\left[-\frac{1}{2}r^T V^{-1} r\right]} \tag{4.40}$$

Considerando que o lado direito da Equação 4.40 é sempre positivo, esta pode ser transformada em:

$$T = 2\ln\lambda = \sup_{b,f_k}\left(r^T V^{-1} r - (r - bf_k)^T V^{-1}(r - bf_k)\right) \tag{4.41}$$

Para a determinação de T se obtém a estimativa de b para qualquer vetor f_k que gera o supremo na equação anterior. Então é calculada a estimativa de máxima verossimilhança:

$$\hat{b} = (f_k^T V^{-1} f_k)^{-1}(f_k^T V^{-1} f_k) \tag{4.42}$$

Substituindo o valor de $\hat{b}$ na Equação 4.41 e chamando o valor correspondente à T de T_k temos:

$$T_k = \frac{{d_k}^2}{C_k} \qquad d_k = f_k^T V^{-1} r \qquad C_k = f_k^T V^{-1} f_k \tag{4.43}$$

Este cálculo é feito para cada vetor f_k do conjunto F e o teste estatístico T é obtido após como,

$$T = \sup_k T_k \tag{4.44}$$

Sob a hipótese H_0, o termo d_k da Equação 4.43 segue uma distribuição normal, com média zero e covariância conhecida C_k. Então, T_k segue uma distribuição qui-quadrado com *um* grau de liberdade sob H_0. Considerando que os valores de T não são independentes, a distribuição de T não pode ser definida. Entretanto, escolhe-se como valor crítico $\chi^2_{(1,1-\alpha^*/2)}$. Para um grau de significância α, o valor de α^* é escolhido como:

$$\alpha^* = 1-(1-\alpha)^{1/p} \tag{4.45}$$

Sendo p o número de erros grosseiros hipotéticos testados (número de componentes do vetor F) e garantindo que a probabilidade de erro tipo I seja menor ou igual à α. Da mesma forma que nos MT e NT, T é comparado com o valor crítico T_c e o erro grosseiro é detectado se $T \geq T_c$. O erro grosseiro que corresponde ao vetor f^* é identificado como erro grosseiro e a sua magnitude é estimada usando f^* para f_k. Na Figura 4.5 é apresentado o fluxograma do algoritmo.

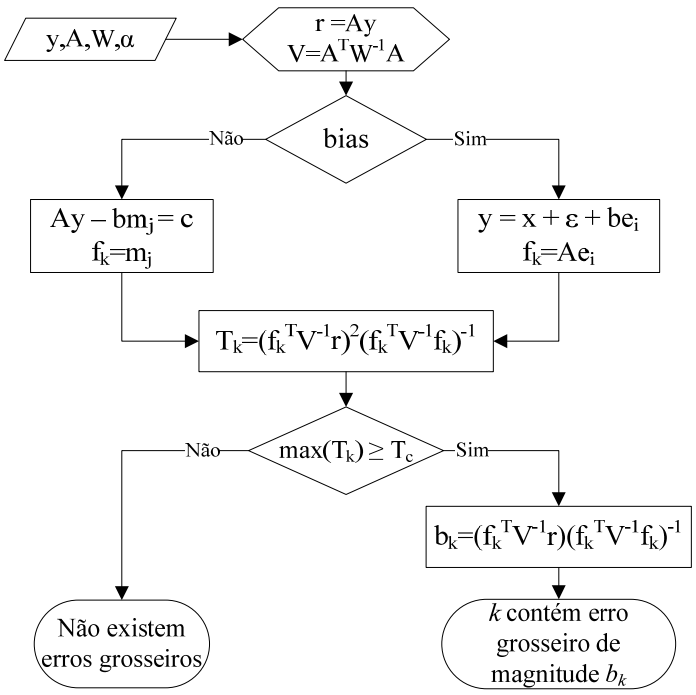

Figura 4.5: Fluxograma do Teste *GLR*.

Nota-se que na formulação apresentada não foi imposta nenhuma restrição que garanta que a magnitude de b deva ser positiva. Narashimhan et al. (1987), sugerem que isto seja feito explicitamente de maneira que, quando a magnitude de b é negativa, esta variável seja descartada como uma possível fonte de erro grosseiro. E da mesma forma pode-se lidar com limites superiores (Narashiman e Mah, 1987).

Vale a pena salientar que o valor numérico do GLR é igual ao valor numérico da raiz quadrada do teste MT. Isto acontece quando se supõe a existência de somente um *bias* na variável i. O vetor de erros grosseiros f_i para um *bias* no sensor i é dado por Ae_i. Assim tem-se:

$$T_i = \frac{\left(e_i^T A^T V^{-1} r\right)^2}{\left(e_i^T A^T V^{-1} A e_i\right)} \tag{4.46}$$

Definindo

$$a^* = A^T (AQA^T)^{-1} r \tag{4.47}$$

$$W_{a^*} = A^T (AQA^T)^{-1} A \tag{4.48}$$

Então, tem-se,

$$T_i = \frac{(a_i^*)^2}{(W_{a^*})_{ii}} \tag{4.49}$$

Caso suponha-se que só existem erros do tipo bias, o lado direito da Equação 4.49 é igual à raiz quadrada do teste da medição (Equação 4.20). Se o mesmo nível de significância for usado para ambos os testes, o critério de teste do GLR é o quadrado do critério usado no teste da medição e a desempenho de ambos os métodos é similar para identificação de medidores com bias (Narashiman e Mah, 1987). A diferença entre os dois métodos é que o GLR pode também identificar vazamentos por uma mudança simples em H_0.

Quando comparado com os outros testes apresentados, na presença de um único erro grosseiro, o teste GLR é o que apresenta maior poder de detecção (Narashiman et al., 2000). Além disto, ainda existe outra versão deste método para inclusão de limites nas variáveis que não será apresentado nesta dissertação. Este método se chama *GLR Restrito* e a principal diferença é a solução do problema de reconciliação por algum método de programação não-linear e a averiguação das restrições ativas do problema de otimização.

4.1.5 Teste da Análise do Componente Principal (PCA)

Este é um teste similar aos testes NT e MT, e é baseado na análise do componente principal (PCA), explorado por Tong el al. (1995). O propósito é redução de tamanho do conjunto de dados pela técnica conhecida como *componentes principais* (PC). Os dados são comprimidos e a informação principal é extraída de forma que os dados são projetados em um subespaço de dimensão menor. Esta transformação visa estabelecer um teste estatístico para as restrições de processo equivalente ao teste NT. Portanto, considerando um conjunto de combinações lineares dos resíduos das restrições r pode-se definir um novo conjunto de variáveis p_r:

$$p_r = W_r^T r = \Lambda_r^{-\frac{1}{2}} U_r r \tag{4.50}$$

Onde a matriz Λ é uma matriz diagonal dos valores característicos de V (a matriz de covariância dos resíduos das restrições). A matriz U_r corresponde aos vetores característicos ortonormalizados de V, de modo que $U_r U_t^T = I$. O vetor p_r tem os componentes principais (PC). Assim, a partir de um conjunto de variáveis correlacionadas r, se gera um novo conjunto de variáveis não correlacionadas p_r e, se as variáveis são normalmente distribuídas e não contêm erros grosseiros, o mesmo acontece com as suas componentes principais (Tong et al., 1998; Romagnoli et al., 2000).

O vetor r pode ser reconstruído a partir das suas componentes principais, mas somente se todas forem retidas. Se somente algumas das n^* componentes principais ($n^* < n_{restrições}$) forem retidas, tem-se que:

$$r = U_r \Lambda^{\frac{1}{2}} p_r + (r - \hat{r}) \tag{4.51}$$

$$\hat{r} = U_r \Lambda^{\frac{1}{2}} p_r \qquad p_r \in \Re^k \; e \; k < m \tag{4.52}$$

Portanto, $\hat{r}$ é o *modelo de componente principal* para r e, a partir da Equação 4.51, pode-se ver que os resíduos do vetor r podem ser decompostos em contribuições a partir das *PCs* e dos resíduos do modelo de componente principal $(r - \hat{r})$. Assim utiliza-se para detecção de erros grosseiros testes de hipóteses em p_r e $(r - \hat{r})$.

Da mesma forma pode-se obter a transformação em componentes principais para os ajustes nas medições visando à construção de um teste estatístico para as medições do processo, equivalente ao teste MT. Considerando um conjunto de combinações lineares dos ajustes nas medições ε.

$$p_a = W_a^T \; a = \Lambda_a^{-\frac{1}{2}} U_a a \tag{4.53}$$

A matriz Λ é uma matriz diagonal dos valores característicos de W_a (Equação 4.17). A matriz U_a corresponde aos vetores característicos ortonormalizados de W_a, de modo que $U_a U_a^T = I$. O vetor p_a tem os componentes principais. Se as variáveis são normalmente distribuídas e não contêm erros grosseiros, então cada elemento do vetor p_a é segue uma distribuição normal padrão com variância W_a. Para o teste de detecção utiliza-se uma regra de detecção similar aos testes *NT* e *MT*. A restrição i ou a medida j será rejeitada e detectada como suspeita de conter erro grosseiro se,

$$p_i \geq Z_{(1-\alpha^*/2)} \qquad \alpha^* = 1 - (1 - \alpha)^{\frac{1}{k}} \tag{4.54}$$

O intervalo de confiança pode ser escolhido como na Equação 4.54, substituindo m pelo número de componentes principais retidas k. A partir da componente principal rejeitada

no teste de hipóteses pode-se identificar a medida, ou restrição, que contém o erro grosseiro avaliando qual é a maior contribuição para a *PC* rejeitada. O fluxograma do algoritmo é apresentado na Figura 4.6.

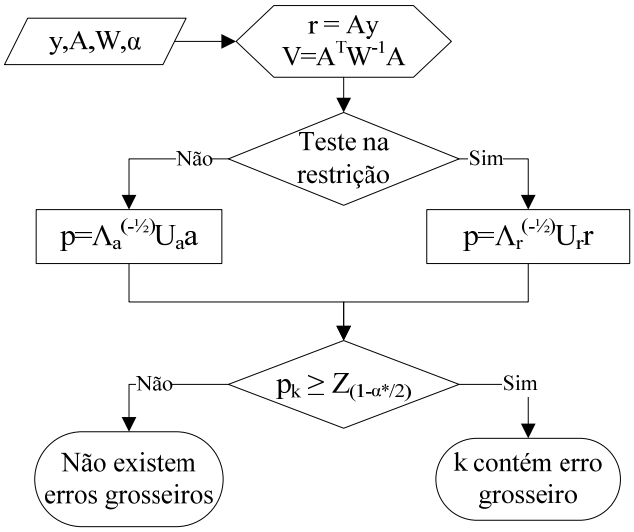

Figura 4.6: Fluxograma do Teste *PCA*.

Apesar deste método utilizar um cálculo mais complexo para a determinação dos vetores e valores característicos, as vantagens de sua utilização, segundo Tong et al. (1995), seriam a remoção da correlação entre as variáveis e o controle da probabilidade do algoritmo cometer erro Tipo I (*falso alarme*). Isto se deve ao fato do teste aplicado ser multivariado, fazendo com que o nível de significância modificado seja menor do que o teste univariado equivalente (por exemplo, MT de máxima Potência). Isto não é necessariamente verdade pois se pode obter o mesmo efeito escolhendo um nível de significância menor (Narashiman e Jordache, 2000). Além disto, os autores indicaram que o teste apresenta um melhor desempenho na detecção de erros persistentes quando comparado com outros testes. Estas afirmações não foram confirmadas no estudo comparativo realizado por Jordache e Tilton (1999).

4.2 Estratégias para detecção de múltiplos erros grosseiros

Por causa do efeito "smearing", não é possível utilizar somente um dos métodos apresentados anteriormente de uma só vez. Quando existem múltiplos erros grosseiros, são utilizadas as estratégias iterativas para detecção de erros grosseiros. As estratégias para a detecção de múltiplos erros grosseiros podem ser seriadas ou simultâneas, dependendo da maneira com que os erros são estimados.

Nas estratégias seriadas, o teste de hipótese é aplicado para todo o conjunto de dados e aquela medição que mais exceder o critério é escolhido como o erro grosseiro. Se o algoritmo for de eliminação seriada, então esta variável é eliminada do conjunto de dados e tratada como uma variável não-medida até que nenhum outro erro seja identificado (Sanchez et al.,

1999). Estas estratégias são simples, com baixo custo computacional, mas apresentam a desvantagem da perda de redundância. Outra desvantagem é que não são aplicados a erros grosseiros que não sejam diretamente associados às medições, como por exemplo, vazamentos (Mah, 1990).

Já se forem aplicadas as estratégias de compensação seriada, o erro grosseiro é estimado e compensado a cada iteração. Desta forma, as estratégias são aplicáveis para todos os tipos de erros grosseiros, podendo manter a redundância durante o procedimento, mas os resultados dependem muito da precisão (acuracidade) da estimação do tamanho do erro grosseiro (Rollins e Davis, 1992).

Nas estratégias simultâneas, é gerada uma lista de candidatos de forma iterativa, mas a compensação dos erros grosseiros ocorre simultaneamente. A fase de identificação de candidatos pode utilizar compensação seriada ou eliminação seriada. E a fase de identificação do conjunto de erros grosseiros pode ser feita de maneira combinatória.

A estratégia mais popular é o *Teste Iterativo da Medição* (IMT), uma estratégia seriada eliminatória, simples e de baixo consumo computacional. O segundo método mais aplicado na literatura da área é a estratégia de *Compensação Seriada Simples* (SSCS, ou ainda a versão modificada, MSCS), seguida pelas estratégias da *Técnica da Combinação Linear* (LCT) e da *Estimação Simultânea de Erros Grosseiros* (SEGE). Estes métodos apresentam uma extensa literatura e por isto são as referências usadas neste trabalho. As outras estratégias (Testes combinados MTNT e NTMT) foram pouco exploradas e, por parecerem interessantes, foram também incluídas. Como são baseados na combinação de testes simples apresentam um grande potencial de poderem ser estendidos para problemas mais complexos.

4.2.1 Teste Iterativo da Medição (IMT)

O teste IMT (“Iterative Measurement Test”) foi proposto por Serth e Henan (1986), usa uma estratégia de eliminação seriada para identificação de bias em medidas e é baseado no teste da medição. O teste é feito iterativamente e a cada iteração, a variável que apresentar a maior estatística Z_a que exceda o valor crítico é declarada como contendo erro grosseiro. Esta é eliminada do conjunto de dados utilizando alguma técnica de redução de tamanho, sendo tratada como se fosse uma variável mão-medida. Na proposta original do algoritmo, os autores utilizam a técnica de agregação nodal, e nesta dissertação é utilizada a decomposição QR.

As etapas, em linhas gerais, são:

1. Preparação: Considere T como o conjunto com todas as variáveis do processo, S contém as medições com erros grosseiros (S é vazio neste momento), R contém das variáveis medidas, U contém as variáveis não medidas e i = n° de variáveis em U.

2. Incrementar contador, $i = i + 1$. Se $i = n$, ir para o passo 7, caso contrário, ir para o passo 3.

3. Reconciliar os dados do conjunto *R*.

4. Aplicar o MT para o conjunto *R*.

5. Escolher a medida com o maior $Z_a \geq Z_{ac}$ e adicionar aos conjuntos *S e U*. Se não existirem $Z_a \geq Z_{ac}$, ir para o passo 7. Caso exista mais de uma medida com $Z_{ai} \geq Z_{ac}$ e com valores iguais, escolher a com menor índice e ir para o próximo passo.

6. Fazer $R = T - U$. Obter *T*, resolver o problema de reconciliação para o conjunto *R* atualizado. Retornar ao passo 2.

7. As medições y_i, $i \in S$ foram detectadas como contendo erros grosseiros. As estimativas reconciliadas após a remoção destas são obtidas no passo 4 da última iteração.

O algoritmo é apresentado em forma de fluxograma na Figura 4.7. Esta é a estratégia mais popular na literatura devido a sua simplicidade e por ser adequada para aplicação em tempo real. As desvantagens são que, devido à utilização de uma referência proveniente de uma etapa de reconciliação intermediária, muitas variáveis podem ser declaradas como contendo erros grosseiros devido ao espalhamento pelos EGs presentes no conjunto de dados. Além disto, só podem ser determinados *n*-1 erros grosseiros e, durante a etapa de estimação, podem ser estimados valores negativos de vazão. Para superar este último problema, foi então proposta uma modificação deste algoritmo, conhecido como *MIMT* que será apresentado a seguir.

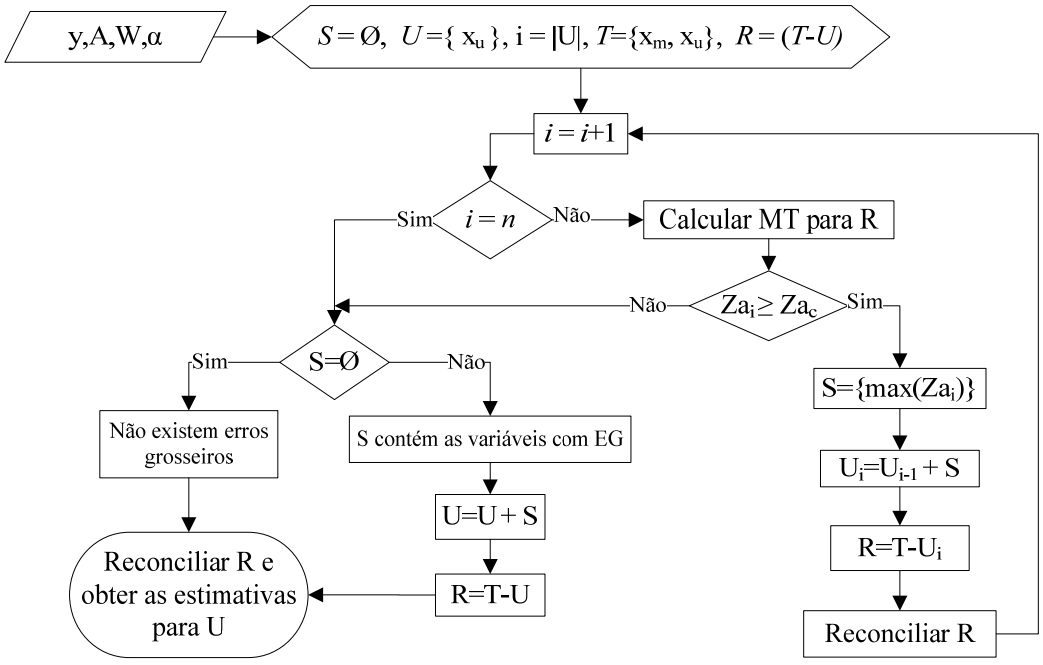

Figura 4.7: Fluxograma do IMT.

4.2.2 Teste Iterativo da Medição Modificado (MIMT)

O MIMT ("Modified Iterative Measurement Test") é uma modificação do *IMT*, proposto por Serth e Henan (1986), para inclusão de limites superiores e inferiores nas variáveis estimadas. A estratégia continua seguindo a filosofia iterativa e eliminatória do IMT, mas a identificação do erro grosseiro não é mais a simples regra de escolher a variável com o maior valor do teste estatístico que supera o valor crítico. Após a identificação da variável candidata, o conjunto de dados é reconciliado e é obtida uma estimativa para o erro grosseiro. Assim, além de exceder o valor critico a estimativa ainda é restrita a valores positivos e a um limite superior.

Em Serth e Heenan (1986), é sugerido que o limite superior seja o valor do topo de escala do instrumento e, para dados provenientes de simulação, este seja de 4 vezes o valor da medida verdadeira. Segundo os autores, esta modificação faz com que o cometimento de erros tipo I seja reduzido em relação à estratégia original. Da mesma forma que ocorre com o IMT, a única modificação em relação ao algoritmo original utilizada nesta dissertação é a utilização da decomposição QR ao invés da técnica de agregação nodal. O algoritmo é praticamente o mesmo que o apresentado para o IMT. O único passo que é modificado é o passo 6. Neste, primeiro verifica-se se as estimativas estão dentro dos limites. Se sim, retorna-se ao passo 2. Caso contrário, volta-se ao passo 5 escolhendo a próxima medida com o maior $Z_a \geq Z_{ac}$. Na Figura 4.8 é apresentado o fluxograma do algoritmo.

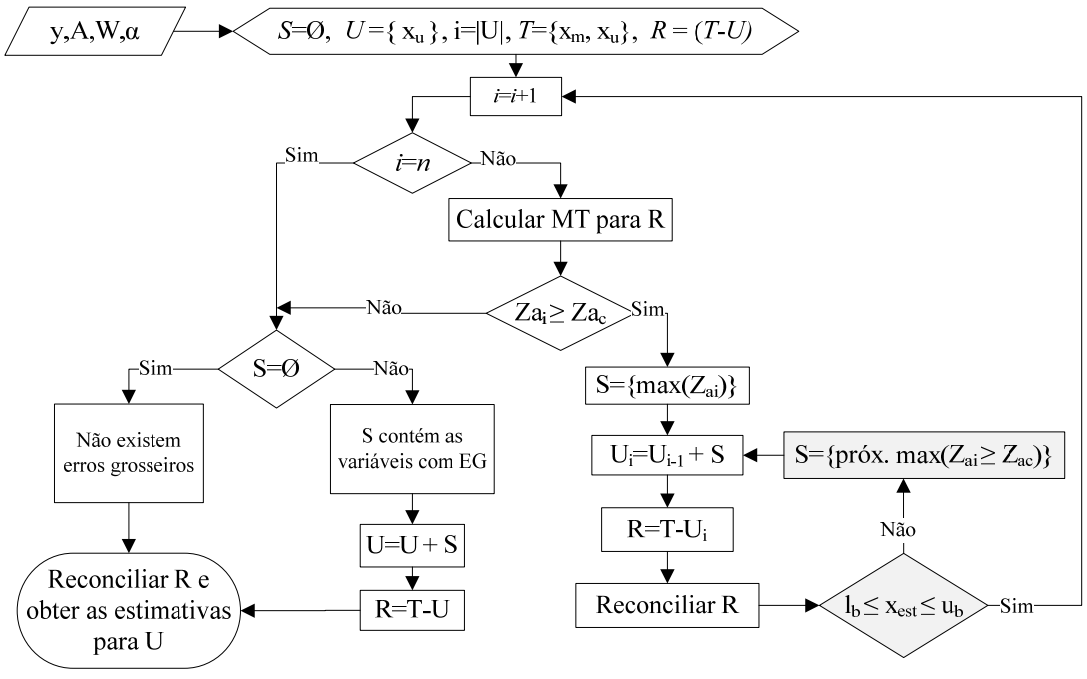

Figura 4.8: Fluxograma do *MIMT*.

Este algoritmo apresenta algumas desvantagens. No final do procedimento, as estimativas não necessariamente irão satisfazer as restrições do problema. O algoritmo pode acabar antes que todos os valores dos testes estatísticos estejam abaixo do valor critico, ou ainda, todas as medidas podem passar no teste estatístico, mas as estimativas não respeitam os limites (isto pode acontecer se, na primeira iteração não são detectados EGs).

4.2.3 Teste Combinado MT-NT

Esta estratégia, sugerida por Yang et al. (1995), é iterativa, eliminatória e combina dois testes já apresentados neste trabalho: MT e NT. Na etapa de detecção e identificação do erro grosseiro, realiza-se uma busca por candidatos avaliando-se os resultados do teste da medição. Para a confirmação é realizada a verificação dos nós relacionados à corrente candidata utilizando o NT. Após este procedimento utiliza-se como critério final o ajuste relativo realizado durante a reconciliação. A estimação do erro grosseiro é feita de modo que a variável detectada é tratada como uma corrente não-medida e seu valor estimado a partir das medições livres de erros disponíveis. O algoritmo é descrito a seguir e o fluxograma é apresentado na Figura 4.9.

1. Preparação: Idem ao IMT

2. $i = i + 1$. Se $i = n$, ir para o passo 10, caso contrário, ir para o passo 3.

3. Aplicar o MT. Se todas as correntes passarem no teste, ir para o passo 10. Caso contrário ir para o próximo passo.

4. Classificar a corrente com o maior $Z_{ai} \geq Z_{ac}$ como $X_{bad.}$ Achar os dois nós que ligados a esta corrente.

5. Aplicar o NT para estes nós. Se o nós passarem no teste da medição, esta corrente não contém erro grosseiro. Voltar para o passo 3 considerando como X_{bad} a corrente com o próximo maior Z_{ai}. Caso contrário considerar os nós como no_{bad} e ir para o passo 6.

6. Achar todas as correntes ligadas aos nós em no_{bad} e calcular os ajustes relativos a_i/y_i.

7. Escolher a corrente com o maior ajuste relativo e aplicar o NT para o outro nó ligado a esta. Se *a corrente não passar no NT,* ir para o passo 8. Caso contrário ir para passo 8.

8. Colocar esta corrente fora do conjunto das correntes suspeitas, e considerar a corrente com o próximo maior ajuste. Voltar para o passo 7.

9. Checar se o ajuste relativo é maior do que 0,05. Se sim, ir para o próximo passo. Caso contrário, X_{bad} contém o erro grosseiro.

10. Mover este erro grosseiro para o conjunto das variáveis não medidas U. Colocar todas as variáveis de S em U. Os dados medidos originais serão usados no conjunto R onde $R = T - U$. Reconciliar R e estimar as variáveis do conjunto U. Voltar para o passo 2.

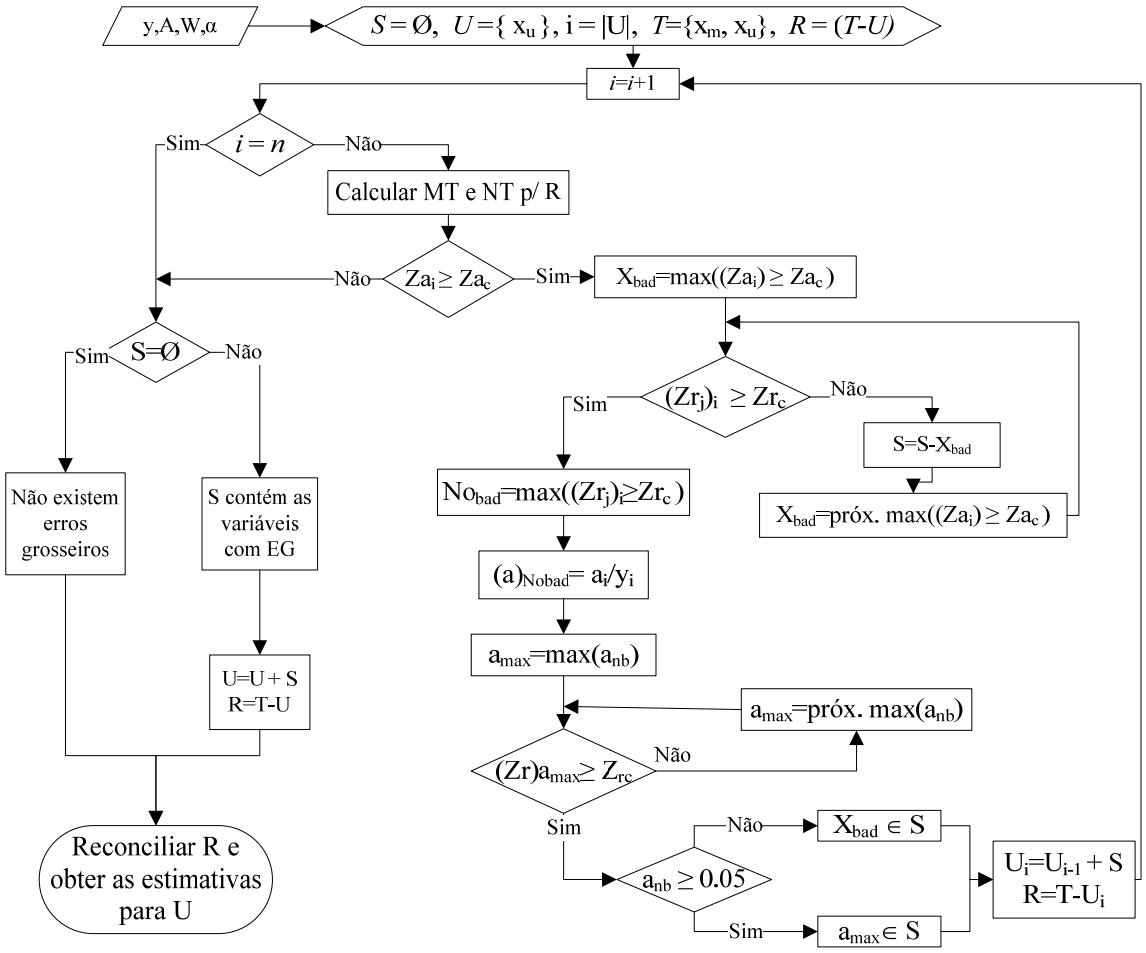

Figura 4.9: Fluxograma do MT-NT.

Este método foi proposto como uma opção para os métodos IMT e MIMT, mas apresenta um cálculo mais complexo. O MT-NT herda as desvantagens dos seus testes bases. Do MT aparece a primeira desvantagem: para criar uma referência, ao se utilizar a solução por mínimos quadrados, o erro grosseiro é espalhado por todos os dados e isto pode causar danos aos dados "bons". Em outras palavras, tende à existência do erro tipo I.

Já o NT passa para a estratégia a segunda principal desvantagem: se existem dois erros grosseiros de mesma magnitude, com sinais diferentes na entrada do nó ou de sinais diferentes um na entrada e outro na saída, eles podem eliminar um ao outro causando uma falha na identificação (Mei et al., 2006). Além disto, não existe maneira de garantir que os erros grosseiros não afetarão todo o conjunto de dados durante a aplicação da reconciliação. Se isto de fato acontecer, o critério do ajuste relativo não refletirá a realidade das medidas com bias, e não deve ser utilizado para identificação das correntes corretamente. (Mei et al., 2006).

Wang et al. (2004) propuseram uma versão compensatória deste algoritmo. Isto diminui o problema de perda de observabilidade devido à diminuição do posto da matriz de coeficientes durante a eliminação seriada das variáveis.

4.2.4 Teste Combinado NT-MT

Este teste, apresentado por Mei et al. (2006), é uma modificação do teste MT-NT. As principais diferenças são que o método utiliza o teste NT para a montagem da lista de variáveis candidatas a erros e o teste MT para a checagem da lista e não utiliza o critério do ajuste relativo para definir qual é a corrente que apresenta o erro grosseiro. Para estimação, utiliza-se uma estratégia de compensação seriada para evitar a diminuição do posto da matriz de coeficientes. Como estratégia para diminuir o efeito do espalhamento dos erros grosseiros, utiliza-se no MT uma matriz de variância unitária. O algoritmo é descrito a seguir, tal qual publicado no trabalho referência. O fluxograma é apresentado na Figura 4.10.

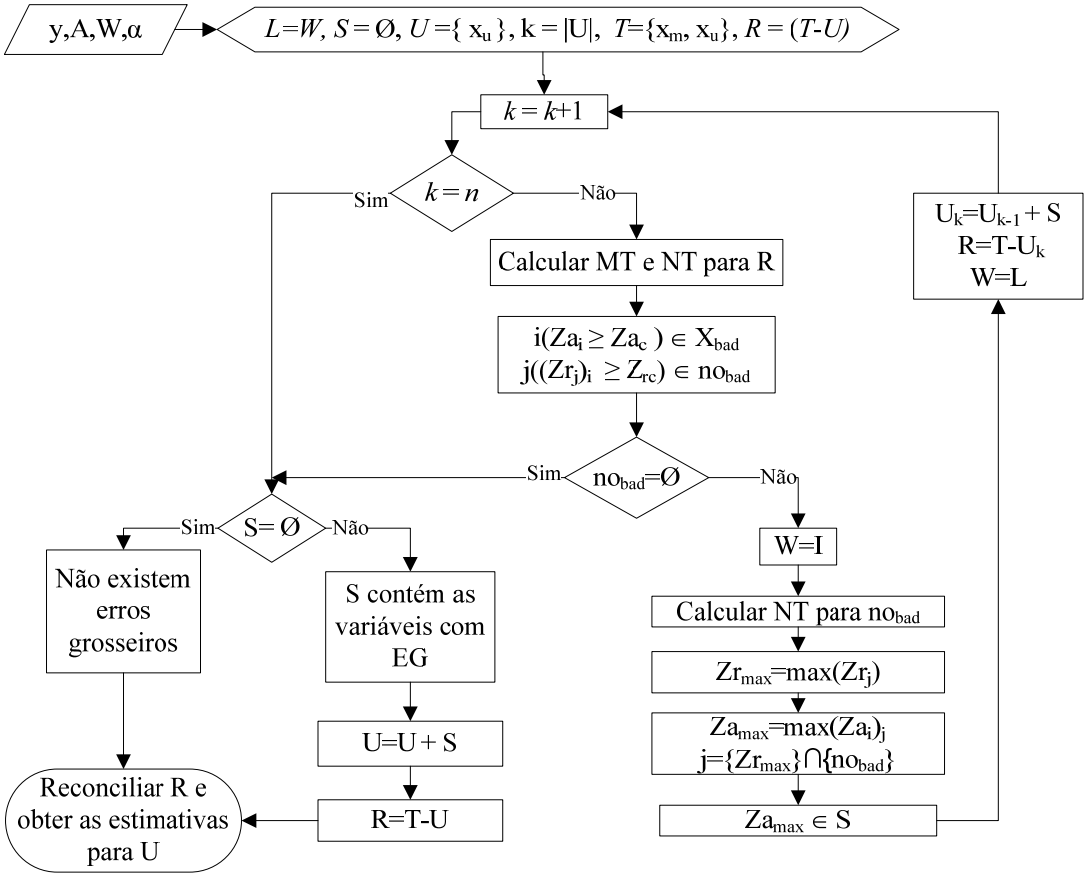

Figura 4.10: Fluxograma do NT-MT.

As etapas são:

1. Aplicar o NT e o MT para *R*.

2. Definir uma nova matriz *L=W*. Os nós identificados pelo *NT* ($Z_{rj} \geq Z_{rc}$) são chamados de no_{bad}. A_s correntes identifica_{das} pelo MT ($Z_{ai} \geq Z_{ac}$) são chamadas de X_{bad}. *Se* $_n o_{bad}$ é um conjunto $_{vaz}$io, ir_{par}a o passo 5.

3. Fazer *W=I* e recalcular o NT novamente e chamar os nós com os maiores valores de Z_{ri} de Z_{rmax}. Encontrar as correntes ligadas á Z_{rmax} que estão em no_{bad} e chamar a corrente com maior Z_a de Z_{amax}.

4. Estimar Z_{amax} e substituir o valor medido pelo valor estimado. Reconciliar o conjunto de dados e substituir os valores medidos pelas estimativas. Fazer W=L e retornar ao passo 1.

5. Reconciliar o conjunto de dados.

As diferenças entre o fluxograma e as etapas descritas se devem à implementação do algoritmo. O trabalho publicado pelo autor é bem sucinto, e não apresenta resultados significativos. Este foi um dos motivos da escolha deste algoritmo, além de parecer interessante para a extensão a problemas mais complexos.

4.2.5 Estratégia de Compensação Seriada Simples (SSCS)

O SSCS ("Simple Serial Compensation Strategy"), proposto por Narashiman e Mah (1987), é similar ao algoritmo IMT, mas pode estimar tanto bias como vazamentos. Usa como base o teste GLR e ao invés de eliminar as variáveis suspeitas, compensa-as. O princípio da compensação das medidas ou do modelo a cada estágio implicitamente considera que os erros grosseiros identificados nos passos anteriores e suas estimativas são corretos.

Considerando que no inicio do estágio k+1, já se tenha identificado k erros grosseiros correspondente ao vetor $[f^*_1, f^*_2, ..., f^*_k]$ e as suas estimativas são dadas por $[b^*_1, b^*_2, ...b^*_k]$. As hipóteses para o estágio *k+1* são:

H_0^{k+1}(somente os erros grosseiros presentes nos estágios anteriores estão presentes)

$$E[r] = \sum_{j=1}^{k} b_j^* f_j^* \quad ou \quad E[r_c] = 0 \tag{4.55}$$

H_1^{k+1}(um erro adicional está presente)

$$E[r] = \sum_{j=1}^{k} b_j^* f_j^* + bf_i \quad ou \quad E[r_c] = bf_i \tag{4.56}$$

Na Figura 4.11 é apresentado o fluxograma do algoritmo. As etapas são:

1. Calcular o resíduo das restrições r e a matriz de covariância do resíduo das restrições V.

2. Aplicar *GLR* conforme a Equação 4.43 para todos os erros grosseiros.

3. Se nenhum erro foi identificado, ir para o passo 6. Caso contrário, o candidato com maior valor do teste contém erro grosseiro. Estimar o erro grosseiro.

4. Compensar o erro grosseiro identificado no passo 3.

5. Retornar ao passo 2, atualizando *r*.

6. Reconciliar o conjunto de dados obtendo $\hat{x}$. Para isto, utilizar as medidas ou (as restrições) compensadas.

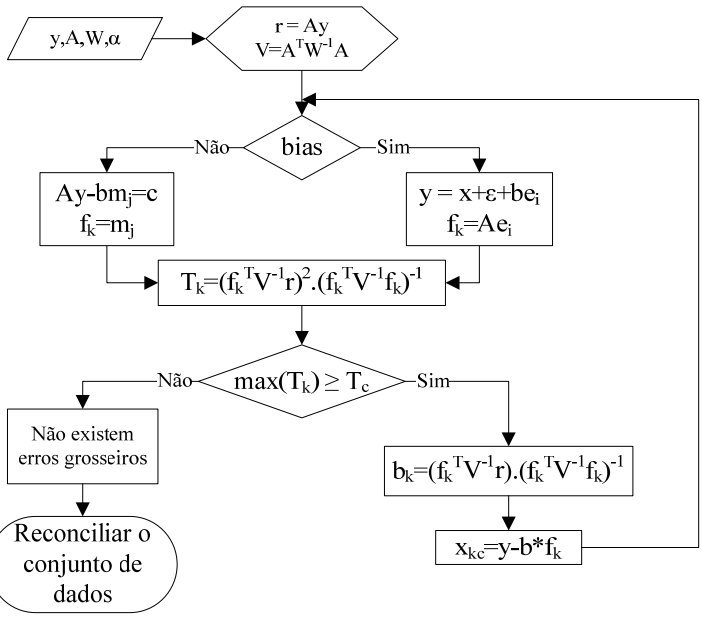

Figura 4.11: Fluxograma do SSCS.

4.2.6 Estimação Simultânea de Erros Grosseiros (SEGE)

O SEGE ("Simultaneous Estimation of Gross Errors") é uma estratégia de compensação simultânea combinatorial proposta por Sanchez e Romagnoli (1994). É composta de um esquema recursivo para isolar os candidatos a fontes de erros grosseiros e então é aplicada uma técnica de identificação e estimação simultâneas para estimação do EG.

O algoritmo é composto de 3 etapas. Na primeira etapa é aplicado o teste GT. Caso o GT acuse a presença de erros grosseiros, é realizada uma etapa de identificação. Para isto, utiliza-se um processamento seqüencial dos dados do processo, onde uma restrição é adicionada por vez e testada frente ao critério. Portanto, na etapa de identificação o objetivo é isolar um subconjunto de restrições que não passam no GT. Para isto adiciona-se uma restrição por vez e aplica-se o teste. Supondo que um problema de reconciliação tenha sido resolvido usando um subconjunto de restrições de processo $G\,x = 0$, onde G é uma matriz (w x z), com $w < m$. Então a matriz de variância para $\hat{x}$ é:

$$W_r^{i-1} = W - WG^T\left(GWG^T\right)^{-1}GW \tag{4.57}$$

Se um conjunto de equações (B_i) é adicionado às restrições, então:

$$\begin{bmatrix} B_i \\ A \end{bmatrix} x = Bx = 0 \tag{4.58}$$

A matriz de variância de $\hat{x}$ para o novo caso é dada pela equação 4.57, substituindo *G* por *B*, ou ainda, em termos da variância do caso anterior $W_r^{(i-1)}$,

$$W_r^i = W_r^{(i-1)} - W_r^{(i-1)} B_i^T \left(B_i W_r^{(i-1)} B_i^T \right)^{-1} B_i W_r^{(i-1)} \tag{4.59}$$

As estimativas e o valor da função objetivo para o novo caso são,

$$\hat{x} = W_r^i W^{-1} y \tag{4.60}$$

$$fobj = (y - \hat{x})^T W^{-1} (y - \hat{x}) \tag{4.61}$$

Após a adição de cada uma das restrições, o valor da função objetivo (*fobj*) é calculado e comparado com o valor crítico. O valor da função objetivo segue uma distribuição qui-quadrado com o número de graus de liberdade *g* = posto(*B)*. Então, para um determinado α, se o valor da função objetivo for maior que o valor crítico, erros grosseiros são detectados e a última equação adicionada é eliminada do sistema. Todas as medidas envolvidas e o nó detectado são adicionados na lista de suspeitos. Caso contrário, erros grosseiros não são detectados, a restrição continua no conjunto e pode-se continuar a adição das restrições até que todas tenham sido testadas. Quando isto acontecer, têm-se a lista de medidas suspeitas.

Após determinar a lista de candidatos, os erros grosseiros são estimados para todas as combinações possíveis dos candidatos. Para isto são testadas todas as combinações possíveis dos candidatos para encontrar a combinação que gera o menor valor de função objetivo do problema de estimação. Inicia-se com *nEG* = 1, obtendo as combinações dos *n* erros. Para um erro do tipo bias, o modelo da medição é modificado para:

$$y = x + \varepsilon B_{rp} m_b \tag{4.62}$$

Onde B_{rb} é uma matriz (*z* x *n*) e m_b é um vetor *s* dimensional com as magnitudes dos erros grosseiros. O problema de mínimos quadrados e a respectiva solução são: ($P_b = AB_{rb}$)

$$\begin{aligned} &\min_{\varepsilon, m_b} \; S = \left\{ \varepsilon^T W^{-1} \varepsilon \right\} \\ &\text{sujeito a : } \; A\left(y - \varepsilon B_{rb} m_b \right) = 0 \end{aligned} \tag{4.63}$$

$$\hat{m}_b = \left[P_b^T \left(AWA^T \right)^{-1} P_b \right]^{-1} P_b^T \left(AWA^T \right) Ay \tag{4.64}$$

$$\hat{\varepsilon} = WA^T\left(AWA^T\right)^{-1}\left[Ay - P_b\hat{m}_b\right] \tag{4.65}$$

$$\hat{x} = y - WA^T\left(AWA^T\right)^{-1}\left[Ay - P_b\hat{m}_b\right] - B_{rb}\hat{m}_b \tag{4.66}$$

Utilizando as Equações 4.64, 4.65 e 4.66 estima-se a magnitude dos erros grosseiros para todas as *c* combinações, e então se calcula os valores das funções objetivo correspondentes e determina-se o valor mínimo. Se, ao aplicar o GT para esta combinação, o teste não detectar EG, então esta é combinação está correta. Caso contrário, o número de EGs é incrementado e todas as novas combinações de nEG + 1 erros grosseiros são testadas. O fluxograma do algoritmo é apresentado na Figura 4.12.

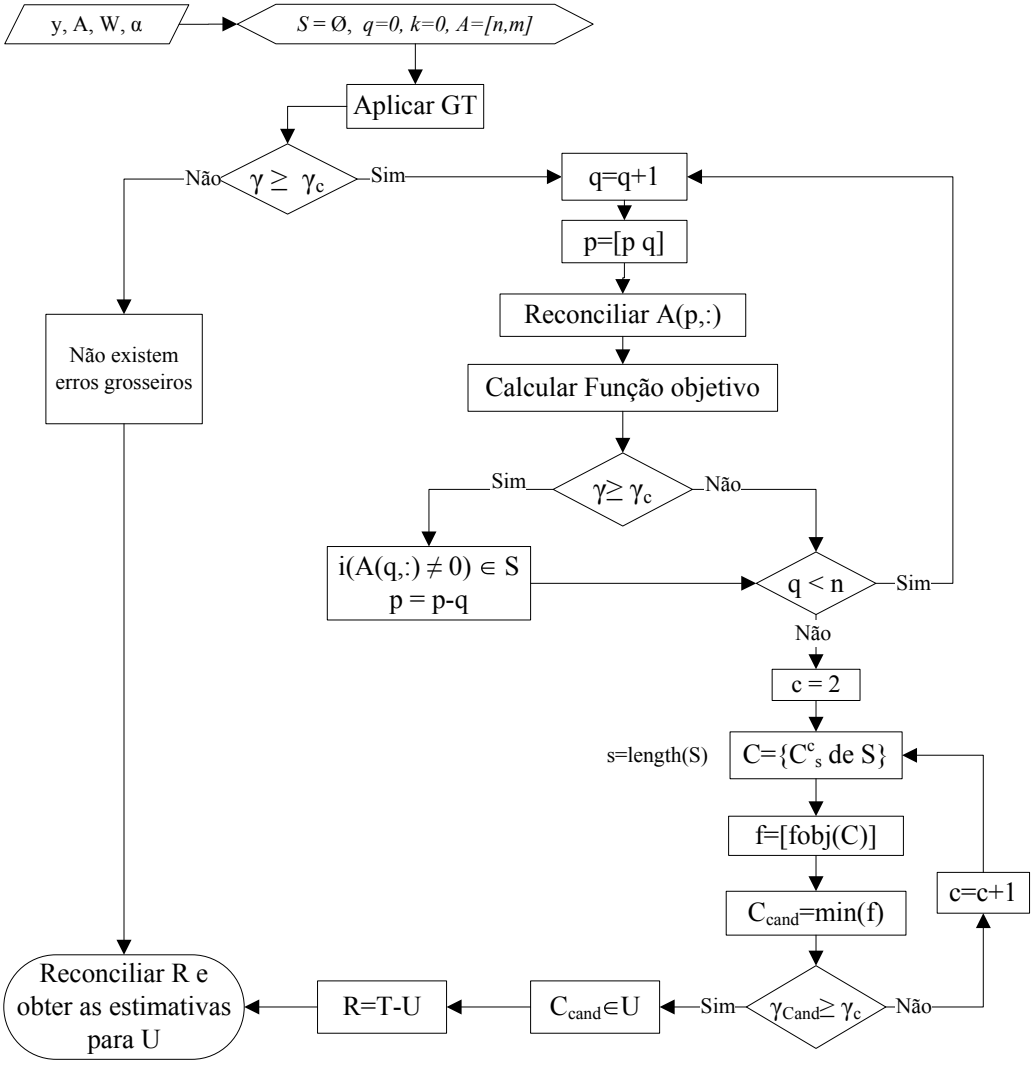

Figura 4.12: Fluxograma do SEGE.

As etapas do algoritmo são:

1. Aplicar o teste global para o conjunto de dados. Se $\gamma \le \chi^2_{n,\alpha}$ não falhar, não existem erros grosseiros e parar. Caso contrário, ir para o passo 2.

2. Adicione uma restrição. Aplicar a reconciliação de dados e determinar o valor da função objetivo.

3. Testar o valor da função objetivo. Se o teste falhar, descartar esta restrição, adicionando as medidas envolvidas na lista de medidas suspeitas. Caso contrário, manter esta restrição no conjunto de teste.

4. Se não existem mais restrições para adicionar, ir para o passo 5. Caso contrário, ir para o passo 2.

5. Definir o número máximo de erros grosseiros hipotéticos (*mnh*).

6. Definir o n° de erros grosseiros k =1.

7. Tomar todas as combinações dos k erros grosseiros, aplicar o modelo da medição e obter o valor da função objetivo.

8. Determinar qual combinação de erros grosseiros gera a menor função objetivo.

9. Aplicar GT para esta combinação. Se $\gamma \geq \chi^2_{m-k,\alpha}$, então esta é a combinação com erros grosseiros. Caso contrário, incrementar k= k+1. Se $k < mnh$ ir para o passo 7.

4.2.7 Técnica da Combinação Linear (LCT)

O LCT ("Linear Combination Technique"), proposto por Rollins e Davis (1992), é baseado no teste NT e testa as combinações entre os nós. A idéia é que se existe um erro grosseiro, este afetará os nós relativos a esta medida e, conseqüentemente, o teste indicará estes dois nós como suspeitos. Se estes nós forem combinados, a corrente que os conecta será eliminada e o pseudo-nó terá maior probabilidade de não ser rejeitado. Assim, ao testar os nós e as suas combinações, as que não forem rejeitadas pelo teste estatístico são assumidas como livre de erros grosseiros.

Em relação às outras estratégias, esta assume que existem tantos erros grosseiros quantos forem possíveis de serem identificados e leva o nome de UBET ("Unbiased Estimation Technique"). Este nome refere-se ao fato dos erros grosseiros serem estimados simultaneamente e, se todos estiverem presentes nos dados, as estimativas serão "unbiased". Rollins e Davis (1992) mostraram que, como este teste considera a presença do máximo possível de erros grosseiros identificáveis, o seu desempenho é melhor quando muitos erros grosseiros estão presentes. Isto difere do esperado para as estratégias que supõem um número mínimo de erros grosseiros (todas as outras apresentadas neste trabalho).

O LCT descreve dois tipos de testes estatísticos para identificar variáveis com bias. O nível de confiança de um dos testes é independente do número de hipóteses testadas (m) e o outro não o é. A estratégia considera primeiro que o número de hipótese é grande e não

definida a priori E, portanto, é independente de m. Esta etapa consiste na avaliação de combinações lineares dos balanços nodais. As hipóteses podem ser descritas como:

$$H_{0i} = l_i^T \mu_r = 0 \tag{4.67}$$

$$H_{1i} = l_i^T \mu_r \neq 0 \tag{4.68}$$

Onde o índice *i* refere-se a i-ésima combinação linear, μ_r é um vetor com os *m* resíduos das restrições e l_i é um vetor de tamanho *n* com 1 na posição que representa a combinação linear dos nós envolvidos no *i-ésimo* teste. Para um balanço com *n* nós existem $2^n - 1$ combinações possíveis (ou testes). O nível de confiança para qualquer valor l_i rejeitará H_{0i} em favor de H_{1i} se e somente se,

$$\frac{N(l^T \bar{r})^2}{(l^T W_r l)} \geq \chi^2_{n,\alpha} \tag{4.69}$$

Onde $\chi^2_{n,\alpha}$ é o (100α)-ésimo percentil superior da distribuição qui-quadrado com *n* graus de liberdade.

E é justamente por causa da Equação 4.69 que a estratégia leva o nome de LCT pois faz uso de uma combinação linear dos resíduos das restrições. Ao contrário das outras estratégias apresentadas, o LCT considera que todas as variáveis medidas no nó em questão possivelmente apresentam bias. Se H_{0i} é rejeitada, a conclusão é de que pelo menos uma das variáveis medidas entrando ou saindo do nó tem bias. Em contraste, se H_{0i} não é rejeitada, a conclusão é de que todas as medições associadas a esta restrição, exceto por aquelas que conectam os nós, não apresentam bias. Diferentemente do primeiro teste, este tipo de teste faz afirmações sobre cada medida.

Após o teste de hipótese, cada variável medida está em um dos dois subconjuntos possíveis. Um subconjunto (C_1), formado pelas variáveis que não contêm bias. O outro subconjunto (C_2) contém as variáveis que *não foi concluído que estão livres de erros grosseiros*. Se as conclusões sobre todos os testes de hipóteses fossem perfeitas e os testes escolhidos fossem suficientes para identificar o máximo possível de variáveis sem bias, os dois subconjuntos teriam as seguintes propriedades: C_1 teria o maior número de medidas sem bias possíveis e C_2 seria o menor possível e conteria todas as variáveis com bias.

Devido à natureza da conclusão para a hipótese nula (pelo menos uma) o subconjunto C_2 pode conter variáveis sem bias. Além disto, como nenhuma conclusão é tirada sobre as variáveis contidas em C_2 então não se cometeu erros do tipo I, até o momento. Erros tipo I serão cometidos mais tarde, quando os parâmetros contidos em C_2 forem estimados. Por outro lado, em relação às variáveis no subconjunto C_1, como a conclusão tirada sobre estas variáveis é que todas estão livres de erros grosseiros, erros do tipo II são cometidos para cada uma das variáveis com bias contidas neste. É importante ter um nível de significância grande porque erros do tipo I em testes de hipótese podem gerar um subconjunto C_2 grande, o que

afetará a acuracidade da estimação. Por outro lado, um poder adequado é importante porque falhar em não rejeitar a hipótese nula falsa colocará variáveis com bias no subconjunto C_1 resultando em erros do tipo II.

Os primeiros dois passos para testar o conjunto de hipóteses são como seguem:

- ✓ Calcular os *n* balanços nodais para todos os nós individualmente.
- ✓ Se H_0 para o nó *k* (*k* = 1, ..., *n*) não for rejeitado, então não realizar combinação com o nó *k*.

A idéia por trás deste critério é que se, por exemplo, $\mu_{ri}=0$, então testar H_0: $\mu_{ri}+\mu_{rj}=0$ é equivalente a testar $\mu_{rj}=0$. E assim, se *s* testes nodais são não rejeitados, o número total de possíveis testes de hipóteses neste ponto é $2^{n-s}-1+s$, que pode ser muito menor que 2^n-1, dependendo de *s*. Testes desnecessários são então eliminados por não se testar as combinações dos seguintes nós:

- Desconectados. Porque ao testar a combinação entre os nós não conectados, é o mesmo que aplicar o teste para os nós individualmente (já realizado na primeira etapa do algoritmo).
- Conectados por uma corrente que já pertença ao C_1 (já classificada como não contendo erros grosseiros): quando uma corrente que conecta os nós foi concluída como não contendo bias, não faz sentido realizar a combinação parar cancelar os efeitos desta corrente porque a conclusão já foi feita – de que não existem bias na medição para ser cancelado.
- Conectados por correntes pequenas: o principal propósito de se testar a combinação entre os nós é eliminar os efeitos das correntes que os conectam. Quando a magnitude da corrente que os conecta é pequena possivelmente não afetará a conclusão sobre a presença de bias nas outras correntes.

O fluxograma do algoritmo é apresentado na Figura 4.13. As etapas são:

1. Considerar C_1 e C_2 conjuntos vazios, calcular *r*, *V* e iniciar contador *i*=0
2. Aplicar o NT.
3. Colocar todas as variáveis relacionadas às restrições onde $Z_{rj} \leq Z_{rc}$ no conjunto S_1, caso contrário colocar as variáveis no conjunto C_2.
4. Combinar os nós onde $Z_{rj} \geq Z_{rc}$ sejam conectados e que a corrente que os conecta não esteja em C_1.

5. Aplicar o NT para as combinações. Se $Z_{rj} \geq Z_{rc}$ então o erro grosseiro não está na corrente que conecta a combinação das restrições e todas as correntes são colocadas no conjunto C_1.

6. As correntes em C_2 foram identificadas como contendo erros grosseiros e serão adicionadas no conjunto *U*. Assim, $R = T - U$ e *R* é o conjunto das variáveis livres de erros grosseiros.

7. Reconciliar o conjunto de dados para *R* e estimar *U*.

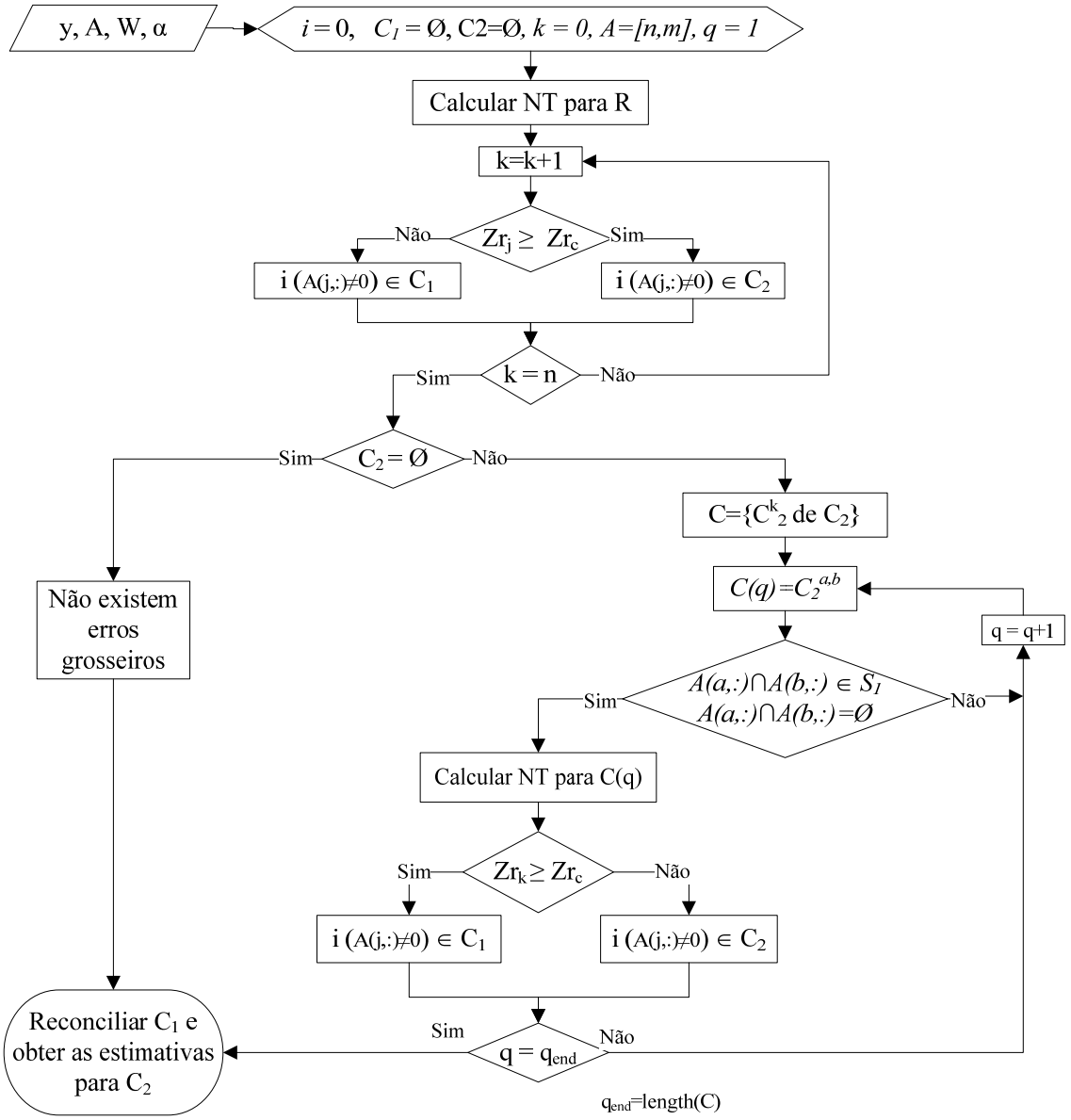

Figura 4.13: Fluxograma do LCT.

A técnica de estimação utilizada nesta dissertação foi a de estimação simultânea utilizada pelo algoritmo GLR. O fluxograma do algoritmo é apresentado na Figura **4.13**. Além desta formulação, esta estratégia ainda pode ser encontrada na literatura utilizando outro teste de detecção. Ao invés de utilizar o teste nodal, Jiang et al. (1999) propuseram uma modificação nesta estratégia utilizando o PCA. Os resultados mostraram que o desempenho da detecção não é significantemente influenciado pela escolha do teste e, sendo assim, a melhor opção é utilizar o NT visto que o cálculo do PCA necessita a obtenção de vetores e valores característicos, o que é mais dispendioso computacionalmente.

4.2.8 Proposta de uma nova estratégia: Teste Iterativo da Medição Robusto (IMT robusto)

Uma das maiores dificuldades das estratégias de detecção é como lidar com o espalhamento dos erros grosseiros nas etapas de estimação intermediárias. Da mesma forma, quando os erros são estimados simultaneamente ao final do procedimento, se a estratégia cometer erros do tipo I, os valores estimados também apresentarão o mesmo problema. Assim, propõe-se nesta dissertação a adequação do método IMT para utilização das funções objetivo robustas, visto que este teste é a estratégia mais popular de detecção de múltiplos erros grosseiros.

A idéia é que, durante a etapa de criação de uma referência para aplicação do MT, ao invés de se utilizar uma função de mínimos quadrados convencional, utilize-se uma função robusta. Como se supõe que para intervalos ao redor da média, a função normal contaminada se aproxima da função normal, o teste de hipótese continua sendo o mesmo. Ao fazer isto, pretende-se evidenciar o ajuste realizado nas variáveis com erros grosseiros.

As etapas do algoritmo são similares as etapas do IMT,

1. Preparação: Idem ao IMT

2. Incrementar contador, $i = i + 1$. Se $i = n$, ir para o passo 7, caso contrário, ir para o passo 3.

3. Reconciliar os dados do conjunto R utilizando a *função objetivo Normal Contaminada.* Calcular o ajuste da medição a = (xest-y).

4. Aplicar o teste da medição para o conjunto R.

5. Escolher a medida com o maior $Z_a \geq Z_{ac}$ e adicionar ao conjunto S, das variáveis contendo o erro grosseiro. Se não existirem $Z_a \geq Z_{ac}$, ir para o passo 7. Caso exista mais de uma medida com $Z_{ai} \geq Z_{ac}$ e com valores iguais, escolher a com menor índice e ir para o próximo passo.

6. Fazer $R = T - U$. Obter T, resolver o problema de reconciliação para o conjunto R atualizado. Retornar ao passo 2.

7. As medições y_i, $i \in S$ foram detectadas como contendo erros grosseiros. Reconciliar as variáveis em R com a função objetivo normal tradicional e estimar as variáveis em S.

Para demonstrar a diferença existente entre o algoritmo IMT tradicional e o algoritmo IMT robusto, será utilizado o caso do trocador de calor com reciclo já apresentado (Figura 2.5, página 21). Neste exemplo, considera-se que existem dois erros grosseiros nas variáveis 2 e 5. Os dados, assim como os resultados, encontram-se na Tabela 4.1.

Tabela

			IMT robusto			IMT		
	y	x	Ajuste	Z_a	xest	Ajuste	Z_a	Xest
1	4.9935	5	-0.05	0.53	5.0067	-0.45	4.83	5.0103
2	20.000	15	4.93	14.34	14.9936	3.47	10.08	18.1924
3	14.9288	15	-0.13	0.40	14.9936	-1.60	4.66	18.1924
4	5.0013	5	-0.00	0.02	4.9941	-0.20	4.25	4.8395
5	14.000	10	3.93	**17.91**	9.9995	2.67	12.17	13.3529
6	5.0000	5	-0.02	0.33	4.9928	-0.88	**13.69**	8.3426
7	5.0271	5	0.01	0.17	5.0067	-0.42	4.48	5.0103

Aplicando o algoritmo IMT tradicional, na primeira iteração, todas as variáveis não passam no teste estatístico, devido ao espalhamento dos erros grosseiros. Diferentemente, na primeira iteração do IMT robusto (utilizando a função objetivo normal contaminada), somente as variáveis 2 e 5 não passam no teste estatístico e, como é esperado, a estimação final é muito próxima do valor verdadeiro. Na aplicação do IMT, as variáveis 5 e 6 são detectadas como contendo erros grosseiros. Desta forma são cometidos erros tipo I e tipo II. Nesta iteração, a única diferença é a maneira com que é calculado o ajuste da reconciliação prévia e este é apresentado na Tabela 4.1.

Nada impede que outras funções objetivo sejam utilizadas neste algoritmo e esta escolha depende da distribuição dos dados e da topologia do processo. No próximo capítulo serão apresentados os testes realizados para a escolha da função objetivo, bem como a comparação de desempenho entre os diferentes estimadores.

Capítulo 5

Metodologia

Para avaliar tanto as estratégias de reconciliação como os algoritmos de detecção, utilizou-se casos amplamente divulgados na literatura e foi realizado um estudo mais abrangente do que o que já existe publicado na área de Reconciliação de Dados. Como resultado deste estudo é proposto uma nova estratégia de detecção de erros grosseiros. Este capítulo tem por objetivo detalhar a metodologia utilizada nesta dissertação, fornecendo detalhes sobre a geração dos dados, simulações e obtenção dos resultados. Todos os algoritmos foram implementados em *Matlab nas versões 5.3 e 7.3*.

De maneira a poder comparar de alguma forma os resultados obtidos com os apresentados pelos mais diversos autores, foram escolhidas as metodologias mais utilizadas na literatura para cada etapa, formando um procedimento de avaliação completo dos algoritmos. Como existem diversos parâmetros que influenciam cada etapa do sistema de reconciliação de dados, foram realizados testes separados para avaliar o desempenho das diferentes etapas:

1) Reconciliação de Dados: Influência da Topologia do processo. Influência do pré-tratamento de dados. Utilização de diferentes funções objetivo.

2) Detecção de erros grosseiros: Sintonia dos métodos de detecção em função do poder de detecção. Influência da topologia do processo. Parâmetros relacionados aos erros grosseiros

3) Teste completo para conjunto detecção-reconciliação: Qualidade da Estimação, Influência da filosofia de detecção nas propriedades de detecção.

Foi realizado um trabalho mais detalhado para um dos estudos de caso de modo que toda a metodologia fosse aplicada. Este escolha se deve à dimensão reduzida do problema, sendo então possível realizar um estudo combinatório completo. Além disto, visto que é um caso muito utilizado, a sua utilização possibilita a comparação com a extensa literatura. Nesta

primeira etapa do trabalho são também descritas as premissas utilizadas para a elaboração da nova estratégia proposta.

Na segunda etapa desta dissertação é proposta a nova estratégia de detecção de erros grosseiros propriamente dita. Esta nova estratégia é validada para o estudo de caso da etapa anterior e ainda reproduz-se um segundo estudo de caso encontrado na literatura.

5.1 Geração dos dados e parâmetros das simulações

5.1.1 Geração do ruído aleatório

Os dados foram gerados utilizando-se a rotina de geração de números pseudo-aleatórios do *Matlab* que recebe o nome de *randn*. Esta função gera uma seqüência de números aleatórios normalmente distribuídos. A partir deste conjunto de dados aleatórios, o ruído aleatório das variáveis foi construído da seguinte forma:

$$\varepsilon = \varpi\sigma x \tag{5.1}$$

onde ω é a variável aleatória, sendo amplificada pelo valor do desvio padrão da variável em questão, representado por $\sigma\, x$. (ex.: 2.5% do valor verdadeiro da variável x_i). Deste modo, tem-se como valor medido da variável i:

$$y_i = x_i + \varepsilon_i = x_i + \varpi_i \sigma_i x_i \tag{5.2}$$

5.1.2 Geração dos erros grosseiros

Os erros grosseiros foram gerados em função do desvio padrão da variável e foi utilizado um algoritmo que leva em consideração 4 fatores:

1) A magnitude (n_σ). Esta foi escolhida em função do número de desvios padrão;

2) O sinal do erro grosseiro (*sig*): Podendo ser + ou –, com mesma probabilidade;

3) A localização do EG (i_{eg}): Sendo que todas as variáveis têm a mesma probabilidade de conterem EG;

4) Número de erros grosseiros (n_{eg}): Escolhido de maneira que ocorressem, no máximo, em 25% do número de variáveis (quando pertinente).

O erro grosseiro é, portanto, gerado da seguinte forma:

$$\delta_{i_{eg}} = (sig)(n_\sigma)(\sigma_{i_{eg}} x_{i_{eg}}), \quad i_{eg} : 1...n_{eg}, \quad n_{eg} \le 0.25n \tag{5.3}$$

Onde δ_{ieg} corresponde ao erro grosseiro adicionado na variável i_{eg}. Portanto, os valores referentes à medição, tanto na presença de erros aleatórios quanto na de erros grosseiros, foram gerados como sendo:

$$y_{i_{eg}} = x_{i_{eg}} + \varepsilon_{i_{eg}} + \delta_{i_{eg}}, \quad i_{eg} : 1...n_{eg}, \quad n_{eg} \le 0.25n \tag{5.4}$$

ou ainda

$$y_{i_{eg}} = x_{i_{eg}} + \varpi_{i_{eg}}\sigma_{i_{eg}}x_{i_{eg}} + sig(n_\sigma)\sigma_{i_{eg}}x_{i_{eg}} \tag{5.5}$$

5.1.3 Simulações Monte Carlo

A utilização de simulações Monte Carlo vem da necessidade de avaliar a probabilidade de detecção dos erros grosseiros. O que se quer é estimar a seguinte integral:

$$\int g(\theta)h(\theta|x)d\theta = E[g(\theta)|x] \tag{5.6}$$

Onde $g(\theta)$ é a probabilidade de θ pertencer à densidade $h(\theta|x)$. Se for possível simular uma amostra aleatória $\theta_1,... \theta_i$ da densidade $h(\theta|x)$, o método Monte Carlo aproxima a integral pela média empírica:

$$\hat{E}[g(\theta)|x] = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \tag{5.7}$$

A qual, pela Lei Forte dos Grandes Números, converge quase certamente para $E[g(\theta)|x]$. A precisão desta aproximação pode ser medida pelo erro padrão (estimado) de Monte Carlo dado por:

$$e_{MC} = \frac{1}{\sqrt{n(n-1)}}\left\{\sum_{i=1}^{n}\left[g(\theta_i) - \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\right]^2\right\}^{1/2} \tag{5.8}$$

Este tipo de método baseado em simulações estocásticas são aproximações, já que a amostra simulada não esgota a distribuição de origem. Contudo o nível de precisão destas aproximações está sob controle do analista que, teoricamente, pode aumentar tanto quanto queira a dimensão da amostra simulada. Os limites práticos impostos pelo custo computacional são cada vez menos severos e, deste modo, os resultados da aplicação deste tipo de método poderão ser eventualmente encarados como quase exatos (Paulino, Turkman, Murteira, 2003).

Todos os resultados que serão mostrados nesta dissertação foram obtidos via simulações Monte Carlo, onde foi utilizado um tamanho populacional constante para o cálculo das probabilidades necessárias para avaliação do desempenho dos métodos. Seguindo

a sugestão de Iordache (1985) cada resultado é baseado em 10.000 rodadas de simulação. Com isto espera-se que a distribuição normal seja bem representada e que os resultados aqui apresentados estejam coerentes com a literatura relacionada. Graficamente, a diferença entre o tamanho da amostra escolhida para simulação pode ser vista na Figura 5.1. Nesta são mostradas 4 amostras de números aleatórios, com tamanhos diferentes, onde a linha contínua é a função distribuição normal correspondente a esta amostra. À medida que o tamanho da amostra aumenta a distribuição normal é mais bem aproximada e o cálculo das probabilidades tende ao valor exato.

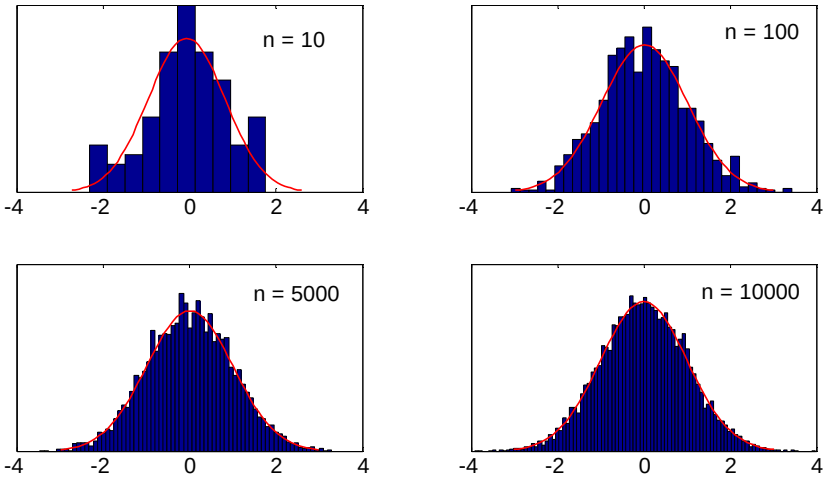

Figura 5.1: Distribuição Normal padrão para diferentes *tamanhos de amostra (n)*.

5.2 Indicadores de desempenho

Nesta dissertação foram utilizados 6 indicadores de desempenho comumente utilizados em reconciliação de dados. Os dois primeiros são relacionados ao problema de estimação propriamente dito e os outros quatro estão relacionados às estratégias de detecção e suas diferentes etapas.

5.2.1 Índice de desempenho para qualidade da estimação

Para quantificar a qualidade global da reconciliação e da detecção utilizou-se o coeficiente de *Redução Total de Erro* (TER) proposto por Serth e Heenan (1987). Este índice reflete a razão entre o ruído restante após a reconciliação e o ruído adicionado para as simulações de todo o conjunto de dados e é dado pela Equação 5.9.

$$TER = \frac{R_{adicionado} - R_{final}}{R_{adicionado}} = \frac{\sqrt{\left(\sum_{i=1}^{m}\left(\frac{(y_i - x_i)}{\sigma_i}\right)^2\right)} - \sqrt{\left(\sum_{i=1}^{m}\left(\frac{(\hat{x}_i - x_i)}{\sigma_i}\right)^2\right)}}{\sqrt{\left(\sum_{i=1}^{m}\left(\frac{(y_i - x_i)}{\sigma_i}\right)^2\right)}} \tag{5.9}$$

Quando o interesse for de avaliar somente o erro de estimação, utilizou-se o índice *ERest*, que reflete a qualidade do valor final estimado e é dada pela Equação 5.10.

$$E\,Re\,st = R_{final} = \sqrt{\left(\sum_{i=1}^{m} \left(\frac{\hat{x}_i - x_i}{\sigma_i} \right)^2 \right)} \tag{5.10}$$

5.2.2 Índices de desempenho para DEG

Para avaliar o desempenho das estratégias de detecção de erros grosseiros são utilizados diferentes critérios relacionados às etapas de detecção, localização e estimação do erro grosseiro.

Para avaliar os algoritmos em relação ao poder de detecção, é utilizado o indicador de *Poder de Detecção Global* (*OP*, *"Overall Power"*) proposto por Narashiman e Mah (1987). Este é dado pela razão entre o número de erros grosseiros corretamente identificados e o número de erros grosseiros simulados, ou ainda,

$$OP = \frac{\sum_{i=1}^{10000} n_{EG}\,corretos}{\sum_{i=1}^{10000} n_{EG}\,simulados} \tag{5.11}$$

Para avaliação da etapa de identificação das estratégias de detecção, é utilizado o indicador *Média de Erros Tipo I* (*AVTI*, *"Average type I error"*), também proposto por Narashiman e Mah (1987) e amplamente utilizado. Este expressa a relação entre o número de erros grosseiros identificados erroneamente pelo número de rodadas de simulação. Este é calculado para cada rodada de simulação separadamente, mesmo que erros grosseiros não sejam simulados. Ou ainda, para esta dissertação,

$$AVTI = \frac{\sum_{i=1}^{10000} n_{GE}\,incorretos}{10000} \tag{5.12}$$

Além disto, ainda utiliza-se o indicador *Fração Esperada de Perfeita Identificação* (*OPF*) proposto por Rollins e Davis (1992) e dado por:

$$OPF = \frac{N°\,de\,rodadas\,perfeitamente\,identificadas}{N°\,de\,rodadas\,de\,simulação} \tag{5.13}$$

Vale a pena ressaltar que após a *Teoria dos Erros Equivalentes* (*TEEq*, proposta por Bagajewicz e Jiang, 1998), estes indicadores foram adaptados para levar em consideração a presença destes conjuntos de erros que não podem ser teoricamente distinguidos. Para isto,

basta que o conjunto de erros grosseiros identificados pertença à mesma classe dos erros grosseiros simulados. Por este motivo, nos resultados de comparação entre as diferentes estratégias, serão apresentadas, sempre que pertinente, 2 colunas. Uma referente ao valor do indicador sem levar em conta a *TEEq* e outra com o valor corrigido.

5.3 Estudo de Caso 1: Rede de trocadores de calor com reciclo (Rosenberg et al., 1986)

Este estudo de caso consiste em uma pequena rede de trocadores de calor com reciclo, proposta por Rosenberg et al., 1986. Na literatura, este caso é também utilizado em Kongjsjahju, Rollins e Bascuñana (2000), Narashiman e Jordache (2000), Romagnoli e Sanchéz (2000), Mei et al. (2006), Narashiman e Mah (1987), Sanchez et al. (1999), Chen et al. (2001), Devanathan et al. (2000), Devanathan et al. (2004), Kao et al. (1990), Rollins e Davis (1992 e 1993), entre outros.

Este é um dos casos mais utilizados na literatura por apresentar uma topologia que gera problemas (os dois reciclos formam ciclo caso não existam algumas variáveis não medidas) e pela facilidade de se realizar estudos combinatórios completos visto que é um sistema pequeno. O fluxograma é apresentado na Figura 5.2.

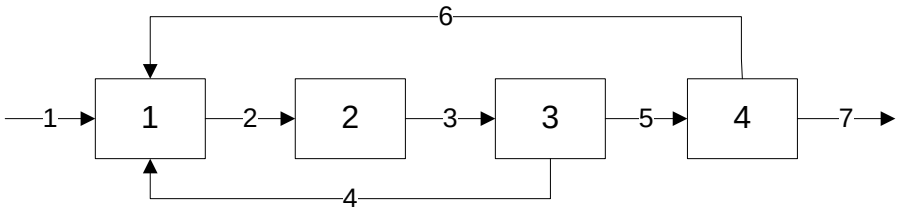

Figura 5.2: Rede de trocadores com reciclo.

A matriz de incidência é dada pela matriz A:

$$A = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix} \tag{5.14}$$

Os dados considerados nas simulações encontram-se na Tabela 5.1. Nesta a segunda coluna corresponde ao valor verdadeiro e σ ao desvio padrão para as variáveis medidas. Este é definido como sendo de 2,5% dos valores verdadeiro da variável (x) e não existe covariância entre as medidas sendo a matriz de variância diagonal. Adicionalmente são mostrados os valores para ajustabilidade e detectabilidade.

Tabela 5.1: Dados para estudo de caso 1.

N°	x	σ	Ajustabilidade	Detectabilidade
1	5	0.13	0.34	0.75
2	15	0.38	0.60	0.92
3	15	0.38	0.60	0.92
4	5	0.13	0.08	0.39
5	10	0.25	0.52	0.87
6	5	0.13	0.14	0.51
7	5	0.13	0.34	0.75

Este caso foi o mais explorado neste trabalho, pois, devido à sua simplicidade, é possível realizar um estudo combinatório completo. As combinações são identificadas pelos números relacionados na Tabela 5.2, sempre que pertinente. Salienta-se que os valores apresentados na Tabela 5.1 referem-se ao caso 0, onde nenhum erro grosseiro está presente e todas as variáveis são medidas.

Tabela 5.2: Combinações utilizadas no caso 1.

Caso	Conjunto	Caso	Conjunto	Caso	Conjunto
1	{1}	22	{3,7}	43	{1,6,7}
2	{2}	23	{4,5}	44	{2,3,4}
3	{3}	24	{4,6}	45	{2,3,5}
4	{4}	25	{4,7}	46	{2,3,6}
5	{5}	26	{5,6}	47	{2,3,7}
6	{6}	27	{5,7}	48	{2,4,5}
7	{7}	28	{6,7}	49	{2,4,6}
8	{1,2}	29	{1,2,3}	50	{2,4,7}
9	{1,3}	30	{1,2,4}	51	{2,5,6}
10	{1,4}	31	{1,2,5}	52	{2,5,7}
11	{1,5}	32	{1,2,6}	53	{2,6,7}
12	{1,6}	33	{1,2,7}	54	{3,4,5}
13	{1,7}	34	{1,3,4}	55	{3,4,6}
14	{2,3}	35	{1,3,5}	56	{3,4,7}
15	{2,4}	36	{1,3,6}	57	{3,5,6}
16	{2,5}	37	{1,3,7}	58	{3,5,7}
17	{2,6}	38	{1,4,5}	59	{3,6,7}
18	{2,7}	39	{1,4,6}	60	{4,5,6}
19	{3,4}	40	{1,4,7}	61	{4,5,7}
20	{3,5}	41	{1,5,6}	62	{4,6,7}
21	{3,6}	42	{1,5,7}	63	{5,6,7}

Neste estudo de caso, a presença de dois reciclos gera um conjunto considerável de erros equivalentes para este sistema. Erros equivalentes acontecem quando dois conjuntos de variáveis que não podem ser, teoricamente, distinguidas um do outro. Para este sistema, existem 3 ciclos, que podem ser obtidos eliminando algumas variáveis (por exemplo, ao utilizar uma estratégia eliminatória). Estes são formados pelas correntes: {2, 3, 4, 5}, {1, 2, 4, 6} e {1, 3, 5, 6}. Qualquer combinação de 3 erros grosseiros dentro de cada conjunto também será equivalente e isto dificulta muito a solução do problema de reconciliação e detecção de erros grosseiros. As combinações são as seguintes:

Classe 1 - {2, 3, 4, 5}: Combinações 44, 45, 48 e 54

Classe 2 - {1, 2, 4, 6}: Combinações 30, 32, 39 e 49

Classe 3 – {1, 3, 5, 6}: Combinações 35, 36, 41 e 57

Portanto, segundo a *TEEq,* sempre que um conjunto de erros grosseiros pertencer à uma destas classes, existe igual probabilidade de que o conjunto correto seja um dos outros pertencentes à mesma classe e estes não podem ser teoricamente detectados. Isto será levado em consideração mais adiante, quando forem obtidos os índices de desempenho dos diferentes algoritmos.

5.4 Avaliação da Reconciliação de dados

5.4.4 Influencia da Topologia

Para avaliar a reconciliação e o efeito da topologia do problema foram simuladas retirada das variáveis em função das combinações possíveis para localização de erros grosseiros. Estas variáveis foram tratadas como não-medidas e obtidas a partir do método da matriz de projeção. Com o valor da variância das estimativas reconciliadas, foi calculada a ajustabilidade (pg. 33).

O objetivo desta etapa é ter uma idéia do que se pode esperar da etapa de reconciliação de dados para os diferentes casos. Inclusive é possível determinar combinações não redundantes na prática, as quais não sofrerão ajuste na etapa de reconciliação.

5.4.5 Influência de pré-tratamento dos dados

A principal diferença entre a filtragem e a reconciliação de dados é que a segunda tem como objetivo reduzir a variância (reduzindo o erro aleatório) garantindo que as restrições de balanço sejam atendidas. A filtragem retira mais os componentes de alta freqüência do sinal, e pode apresentar algumas desvantagens. Existe uma série de problemas relacionados com o projeto de filtros e com a sua sintonia ("*overshooting"*, atraso no tempo,...).

Segundo Narashiman e Jordache. (2000), cada tipo de filtro tem vantagens e desvantagens. Alguns têm a habilidade de produzir significante redução no ruído, mas introduzem um grande atraso na resposta filtrada. Em contra partida existem os que não introduzem tanto atraso, mas também não produzem uma redução no ruído satisfatória. Já outros tipos conseguem remover o ruído sem introduzir grandes atrasos, mas apresentam um desempenho ruim para medições com freqüência variável e o "*overshooting"* é um problema comum nestes casos.

O principal objetivo da utilização de técnicas de filtragem para pré-tratamento de dados é o de melhorar o desempenho dos algoritmos de detecção, pois não deixam de ser técnicas de redução de variância e são freqüentemente utilizados nos Sistemas de Controle Digital Distribuído (SDCD). Para isto foram testados 3 filtros diferentes. O primeiro é um filtro do

tipo FIR passa-baixa, utilizando a implementação *filt* do *Matlab*. O segundo filtro testado é do tipo IIR passa-baixa, onde usou-se a rotina *cheby1*. Por fim testou-se um suavizador do tipo Savizky-Golay, utilizando a implementação *svgolay*.

Estes 3 tipos de filtros foram escolhidos por realizarem operações lineares nos dados, não comprometendo a distribuição final. Demonstrações de provas teóricas e os equacionamentos dos filtros podem ser encontrados em Lathi (1998).

Desta forma, sem entrar na parte teórica de filtragem propriamente dita, a idéia é verificar se a utilização destes tem algum efeito (tanto positivo quanto negativo) sobre a reconciliação e a detecção de erros grosseiros, visto que é um pouco intuitivo pensar que a filtragem poderia atrapalhar a reconciliação já que mudaria as características do ruído. Os filtros tipo FIR e IIR foram escolhidos por serem bem populares. Já o filtro de Savitz-Golay foi escolhido por ser conceitualmente diferente dos outros dois. Este é um suavizador utilizado quando a forma do sinal não é conhecida.

Os filtros tipo FIR (*Finite Impulse Response*) são filtros do tipo média móvel, onde o efeito de qualquer entrada só é sentido por N janelas. Todos os dados de entrada recebem o mesmo peso $w = (1/N)$. Este é considerado um filtro mais eficiente para estimar o ponto central do que o valor atual da variável, e por isto, mais adequado para estimar valores fixos ou tendências lineares de dados (caso do estado estacionário). Já os filtros do tipo IIR (*Infinite Impulse Response*) tem a propriedade de manter o efeito de qualquer entrada infinitamente mas de maneira diminuída. Não apresenta *overshoot* e gera uma boa aproximação para o estado estacionário. Por este motivo é utilizado em muitos *SDCD*s. A sua maior desvantagem é o atraso no tempo gerado quando o ruído é muito atenuado (Narashiman et Jordache., 2000).

O filtro do tipo Savitzky-Golay (*digital smoothing polynomial filter ou least-squares smoothing filters*) é um suavizador tipicamente utilizado para suavizar um sinal com ruído no qual a distribuição da freqüência é grande. Nesta aplicação, o suavizador *Savitzky-Golay* tem um desempenho muito superior aos filtros do tipo media móvel *FIR*, os quais tendem à filtrar uma parte significante do ruído de alta freqüência. Apesar de serem mais eficientes em preservar os componentes de alta freqüência, apresentam desempenho inferior aos filtros FIR em rejeitar ruído. Este é um filtro ótimo no sentido de minimizar o erro quadrático ao se ajustar polinômios à janela de dados (*Mathworks*, 1998).

O principal objetivo dos testes foi comparar o desempenho para redução de variância nos dados em 4 condições: após a filtragem, após a filtragem seguida de reconciliação e após as seqüências filtragem-média-reconciliação e média-filtragem-reconciliação. Os resultados foram comparados com a reconciliação dos valores brutos ou a média destes, dependendo do caso.

Os filtros foram projetados utilizando duas premissas: que todos apresentassem a mesma redução de erro total (TER em torno de 80%) e que não gerassem correlação temporal muito severa no conjunto de dados (o que não seria um problema para a reconciliação de dados estacionária mas no caso de aplicação da técnica na forma dinâmica). Esta segunda

premissa foi monitorada pela avaliação do espectro de potência do sinal resultante, assim como a retirada das altas freqüências obtida pela filtragem. Na Figura 5.3 é apresentado o espectro de potência dos 3 filtros implementados e para o ruído gaussiano, só para comparação. Pode-se verificar que os três filtros atenuaram as altas freqüências, associadas ao erro aleatório inerente da medição.

Foi gerado então um conjunto com 10.000 valores aleatórios para cada uma das variáveis e cada conjunto foi tratado como se fosse um histórico de dados onde foram aplicados os filtros. Após esta etapa, os dados foram separados em 10.000 conjuntos de dados do processo, com todas as variáveis medidas e reconciliados utilizando a solução analítica.

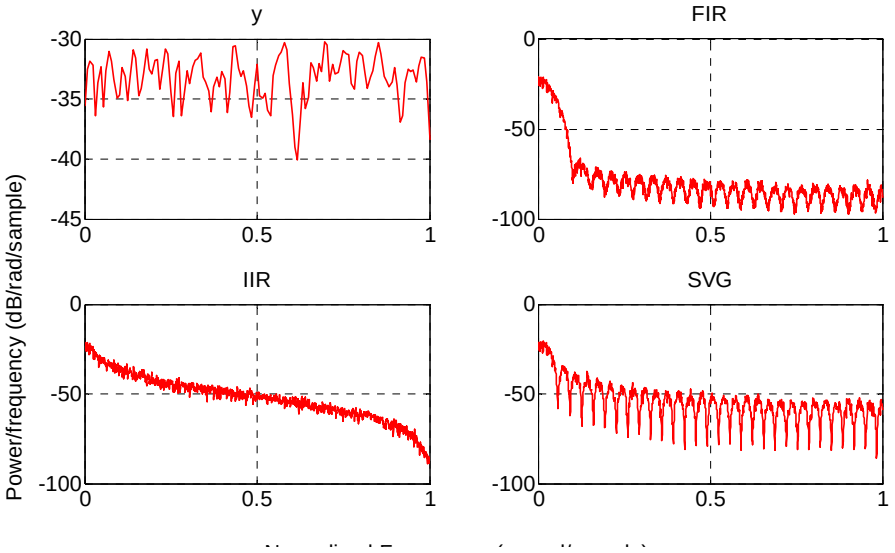

Figura 5.3: Diagramas de espectro de potência para os diferentes filtros implementados.

5.5 Avaliação das estratégias de DEG

Em linhas gerais, para avaliação dos métodos de detecção de erros grosseiros foram realizadas comparações do desempenho obtido nas etapas de detecção, identificação e estimação dos erros grosseiros. Para isto utilizou-se um estudo de caso (Estudo de caso 1) e avaliou-se o efeito da topologia, do tamanho do erro grosseiro e do nível de confiança utilizado nos testes estatísticos. As estratégias foram então sintonizadas para que pudessem ser comparadas na mesma base e avaliou-se o efeito do pré-tratamento dos dados. Finalmente os resultados foram comparados com os obtidos utilizando-se técnicas de reconciliação robusta nas mesmas condições.

Para avaliar a topologia do estudo de caso e o comportamento do efeito *smearing* utilizou-se o conceito de observabilidade. Foram realizadas simulações para todas as combinações possíveis de erros grosseiros e avaliados os resultados. As simulações utilizadas foram obtidas de forma determinística para facilitar o entendimento. Estas são equivalentes à variância das medidas iguais a zero (ou ainda, a não existência de erros aleatórios nas medidas) e servem para definir os conjuntos de erros equivalentes presentes no sistema, caso

não seja óbvio (Rollins et al., 1998). Este estudo é baseado em estudo equivalente realizado por Narashiman e Jordache. (2000).

Em um segundo momento, partiu-se para a comparação entre as diferentes estratégias de detecção de erros grosseiros e esta avaliação foi feita com base em 3 tipos de resultados. O primeiro tipo foi obtido por simulações determinísticas para todos os casos possíveis para o Estudo de caso 1. Este é um estudo combinatório completo, onde foram obtidos os melhores resultados possíveis de desempenho das estratégias de detecção. O segundo tipo de avaliação feita foi com base no levantamento das curvas de poder de detecção em função do tamanho do erro grosseiro e do intervalo de confiança. Com base nestas curvas pode-se notar a necessidade de se encontrar uma forma justa de comparação entre os métodos. Por isto todos os algoritmos foram sintonizados.

Esta forma de comparação foi sugerida por Rosenberg et al. (1986) e resolve o problema da dependência do desempenho das estratégias com a definição dos intervalos de confiança. Esta dependência está relacionada ao fato de que, sempre é possível aumentar o poder de detecção dos métodos de detecção, mas a quantidade de falsos-alarmes aumentará também e por isto, é interessante que os algoritmos sejam comparados na mesma base (Narashiman e Jordache, 2000; Romagnoli e Sánchez, 2000; Rollins et al., 1996). Esta premissa para a comparação justa entre os métodos é utilizada em alguns trabalhos publicados na área (Rosenberg et al., 1987; Rollins et al., 1996; Sanchez e Romagnoli; 1999; Romagnoli e Sanchez, 2000; Narashiman e Jordache, 2000; Jiang et al., 1999; Chen et al., 1998).

Por fim foi realizada a comparação do desempenho das diferentes técnicas com a utilização dos métodos de reconciliação robusta. Quando possível os resultados foram comparados com os encontrados na literatura, mesmo que não tenham sido encontrados trabalhos que relacionassem um número tão grande de estratégias diferentes como o que está sendo apresentado nesta dissertação. Para quantificar os resultados, foram utilizados os indicadores de desempenho amplamente divulgados na literatura.

5.5.1 Simulações determinísticas

A *solução determinística* para o problema de detecção é o melhor que esta aproximação pode fazer. Ela equivale a dizer que os dados obtidos são provenientes de uma média quando $n \rightarrow \infty$, ou ainda, quando não existe erro aleatório ($\varepsilon = 0$). Se pudéssemos medir as variáveis perfeitamente e não fosse possível o cancelamento dos erros, então o resultado determinístico seria garantido testando-se todas as hipóteses possíveis (Rollins et al., 1996).

Para verificar o poder de detecção determinístico das estratégias, estas foram submetidos às simulações sem erro aleatório e foram testadas todas as combinações possíveis, para a presença de um, dois e três erros grosseiros. Quando na presença de 3 erros grosseiros, existe a problemática da teoria dos erros equivalentes. Desta forma serão apresentados os resultados, sempre que pertinente, em duas colunas: a primeira sem levar em consideração esta teoria e a segunda levando-a em consideração. Com base nestes resultados já se pode ter uma idéia de quais algoritmos apresentam alto poder de detecção. Para deixar o

comportamento mais evidente os erros grosseiros foram escolhidos como sendo 10 vezes o desvio padrão.

5.5.2 Levantamento das curvas de poder de detecção

As curvas de poder dos algoritmos foram geradas, de modo que seja possível definir o comportamento dos algoritmos em função do tamanho do erro grosseiro. Poucas destas curvas são apresentadas na literatura, o que impede a comparação. Além disto, alguns algoritmos (MTNT e NTMT) não apresentam dados provenientes de simulações Monte Carlo e foram publicados com base em resultados de uma única realização.

Primeiro foram obtidas as curvas em função do tamanho do erro grosseiro para um nível de confiança constante e na presença de um erro grosseiro. Esta metodologia é baseada no trabalho publicado por Devanathan et al. (2004). Foi escolhido o nível de confiança de 0,05 (95%) pois é o valor encontrado com mais freqüência na literatura. Foram utilizadas simulações Monte Carlo de maneira que o conjunto de erros aleatórios das variáveis fosse constante, mas a localização dos erros grosseiros, bem como a sua magnitude fosse variada. Nesta etapa, para cada rodada foi adicionado um erro grosseiro em cada variável do processo, de um tamanho definido em função do desvio padrão (i.e., a primeira rodada com um erro grosseiro de um desvio padrão em cada uma das variáveis, uma por vez) e este conjunto de dados foi sujeito à todas as estratégias de detecção. Variou-se o tamanho do erro grosseiro de 1 - 40 desvios padrões e o sinal do erro foi escolhido de maneira aleatória com igual probabilidade.

Em um segundo momento levantou-se as curvas de poder de detecção em função também do intervalo de confiança. Para isto foram utilizada simulações nos mesmos moldes das já apresentadas mas o nível de confiança foi variado de 0,01 a 0,4. Para cada combinação de desvio padrão e nível de confiança, foram realizadas 10.000 × (n° de variáveis) × (n° de níveis de confiança) × (n° de desvios padrões) simulações. Assim, no caso do primeiro estudo de caso foram realizados $10.000 \times 7 \times 37 \times 40 = 103.600.000$ simulações para levantar cada curva de poder para cada algoritmo de detecção.

5.5.3 Determinação de intervalos de confiança

Devido à natureza de alguns dos algoritmos de detecção de erros grosseiros, não é possível prever o comportamento da probabilidade de erros tipo I simplesmente especificando o nível de confiança (α). Muitos algoritmos utilizam múltiplos testes simultâneos e, em cada etapa, varia-se o número de testes realizados. Conseqüentemente, são utilizados diferentes valores críticos, escolhidos em função do número de testes estatísticos sendo realizados e para que possam ser comparados é necessário que o valor de α seja determinado por tentativa e erro (Rollins et al., 1996).

Assim, o nível de confiança foi escolhido de maneira que apresente a média de erros tipo I (AVTI) igual 0,1 sob hipótese nula (quando não existem erros grosseiros). Este critério foi proposto por Rosenberg et al., (1986) e é reportada a sua utilização em alguns trabalhos mais recentes como sendo a maneira mais justa de comparar as estratégias (Rollins et al.,

1996; Romagnoli et al., 2000; Narashiman et al., 2000; Sanchez et al., 1999, Chen et al., 1996 e 1998). Para isto foram realizadas 10.000 rodadas, considerando somente a presença de erros aleatórios e o valor de α foi variado até que fosse atingido o patamar de AVTI = 0,1. Uma metodologia análoga a esta é descrita em Chen et al. (1996 e 1998) para ajustar o tamanho mínimo de *bias* a ser detectado quando os algoritmos são aplicados em dados reais.

5.5.4 Simulações Monte Carlo

Após a sintonia dos algoritmos, foram realizadas simulações Monte Carlo de maneira similar às realizadas de forma determinística. Para cada combinação de erro grosseiro foram obtidas 10.000 rodadas para avaliar o desempenho da detecção na presença do ruído.

Foram testadas todas as combinações possíveis para 1, 2 e 3 erros grosseiros. O tamanho do erro grosseiro foi escolhido aleatoriamente entre 5-40 desvios padrões. O sinal do erro grosseiro também foi escolhido da mesma forma, podendo ser positivo ou negativo com igual probabilidade. Quando possível os resultados foram comparados a dados apresentados na literatura.

5.5.5 Influência do pré-tratamento de dados

Muitos autores utilizam dados médios para realizar os testes nos conjuntos de dados. Utilizar as médias é o mesmo que reduzir a variância dos dados medidos (Rollins et al., 1996) e é necessário que seja realizada a correção da variância (o que a maioria das vezes não é explicitado nos artigos encontrados na literatura). Obviamente quando os dados testados apresentam uma variância menor do que a definida na função objetivo aumenta o poder de detecção dos algoritmos e em contrapartida, *bias* com pequenas magnitudes não serão detectados. Em compensação, se a variância dos dados for maior, o AVTI aumenta pois os algoritmos detectarão inclusive alguns dos erros aleatórios. É muito comum na literatura da área a utilização de uma janela para a média igual a 10.

O esperado é que, da mesma maneira que acontece com a média, o pré-tratamento utilizando filtragem dê bons resultados. A idéia de se utilizar o pré-tratamento nesta etapa é verificar o comportamento dos algortimos com dados mais próximos dos reais encontrados na indústria. Muitos dados de processo já vem filtrados e por isto se quer saber se a filtragem traz algum benefício. Para comparar o desempenho dos algoritmos, foram aplicados os 3 filtros já utilizados e os dados foram submetidos aos métodos de detecção já sintonizados para os casos 1 ao 7. Os erros grosseiros foram gerados de maneira similar às simulações Monte Carlo apresentadas no item anterior. Não foram encontradas referências na literatura para comparação. Os únicos trabalhos neste sentido exploram a utilização de médias.

5.5.6 Comparação das estratégias de detecção com Reconciliação Robusta

Para comparar as estratégias de detecção tradicionais com a reconciliaão robusta foram utilizadas simulações determinísticas, nos mesmos moldes das já apresentadas. Para ser possível a comparação entre as diferentes funções objetivo é necessário que estas sejam

sintonizadas para a determinação das constantes. Como é interessante comparar os resultados com os publicados na literatura, as constantes de sintonia das funções objetivo foram utilizadas de acordo com as publicadas em Ozyürt e Pike (2004). Estes valores são apresentados na Tabela 5.3.

Tabela 5.3: Constantes de sintonia para Estimadores Robustos.

Função	Constantes de sintonia
Normal Contaminada (CN)	b = 10 p = 0,235
Cauchy (CA)	c = 2,4
Fair (FA)	c = 1,4

Estas constantes foram obtidas a partir de simulações Monte Carlo de maneira que a variância obtida para os valores estimados fosse de 95% da variância que seria obtida se fosse utilizada a função objetivo normal. Este procedimento foi proposto por Hampel (2002). Como na reconciliação robusta não é realizada a etapa de detecção propriamente dita, comparou-se o resultado final da estimação do conjunto de dados representados pelo índice TER. Para comparação entre os diferentes estimadores robusto foi realizada a sintonia.

5.6 Desenvolvimento e validação da nova estratégia de detecção de erros grosseiros - *IMT robusto*

O *IMT robusto* baseia-se na utilização do algoritmo IMT tradicional, aliado à obtenção da solução do problema de otimização utilizando uma função objetivo robusta. Teoricamente, ao se utilizar uma função objetivo robusta, as estimativas para os ajustes em uma reconciliação prévia (utilizadas no IMT) apresentarão menor efeito "smearing", fazendo com que o indicador de AVTI (*falso alarme*) seja reduzido. Se o AVTI é menor, espera-se que as etapas de detecção e a localização dos erros grosseiros apresentem melhor desempenho. Além disto, se os erros detectados são *todos os erros grosseiros existentes* no conjunto de dados, então a etapa de estimação será "unbiased" (propriedade da estimação por máxima verossimilhança) e a estimação também deverá apresentar melhores resultados.

Neste item serão apresentados os testes realizados para a obtenção da estratégia proposta. Para isto, foi realizada uma etapa de desenvolvimento, comparando entre diferentes tipos de função objetivo e duas maneiras distintas de filosofia de detecção. Estas são também comparadas com o método tradicional. Em um segundo momento é realizada a validação do método, onde são comparados resultados para diferentes estudos de caso existentes na literatura, com os resultados obtidos com a nova estratégia proposta.

5.6.1 Desenvolvimento do método: simulações determinísticas

Para o desenvolvimento do IMT robusto, partiu-se da idéia de investir esforços no método que parecesse mais promissor, já que um estudo baseado em simulações estocásticas demanda muito tempo computacional. Assim, foram realizadas simulações determinísticas

para definir a função objetivo mais eficiente e a filosofia de detecção adequada. Este estudo foi realizado com as mesmas condições já apresentadas na seção 5.5.1.

Como parte deste estudo foram testadas duas filosofias diferentes para a implementação do método. Na primeira, chamada de *IMTr1,* foi implementada a filosofia iterativa do IMT - um erro grosseiro é identificado por vez, a variável é tratada como não medida, e o seu valor é estimado. Já no algoritmo *IMTr2* foi implementada a eliminação de todas as variáveis que não passaram no teste da medição. Se o teste da medição, na versão robusta, conseguir identificar todos os erros grosseiros existentes de maneira correta (e realmente tem maior chance de acontecer, pois o efeito "smearing" é diminuído pela utilização da função objetivo robusta), então a etapa de estimação é "unbiased" e esta é a melhor estimação possível. Se forem tratados um erro por vez, em algumas iterações existirão erros grosseiros e a detecção do próximo candidato pode ser prejudicada. A desvantagem é que deve ser definido um número máximo de erros grosseiros que o processo pode conter. Isto é necessário para garantir que o sistema seja estimável (posto da matriz reduzida das restrições deve ser maior do que o número de variáveis a serem eliminadas do conjunto). Esta restrição aparece devido à falta de consistência estatística do Teste da Medição, demonstrado em Bagajewicz (2005).

5.6.2 Validação do IMT robusto

Para validar o método foram utilizados dois estudos de caso. O primeiro já foi apresentado. Os resultados das simulações Monte Carlo obtidas para todos os métodos de detecção foram comparados com os obtidos pelo novo método.O segundo estudo de caso é amplamente divulgado na literatura e os resultados foram comparados com os trabalhos já publicados. Para este foi escolhido um trabalho de referência, em que a metodologia de teste foi repetida visando a comparação e validação do novo método. O segundo estudo de caso será apresentado na próxima seção.

5.7 Estudo de Caso 2: Rede de vapor

Este estudo de caso consiste em uma rede de medição de vapor de uma unidade de síntese de metanol proposta por Serth e Heenan, 1986. Na literatura são encontradas duas versões do mesmo estudo de caso. Uma com 28 correntes e outra com 25. A segunda versão é a utilizada nesta dissertação e pode ser obtida a partir da primeira por agregação nodal. Da mesma forma que no caso anterior, o desvio padrão das variáveis medidas é considerado de 2,5% do valor verdadeiro (x) e os dados utilizados podem ser encontrados na Tabela 5.4. O fluxograma é apresentado na Figura 5.4.

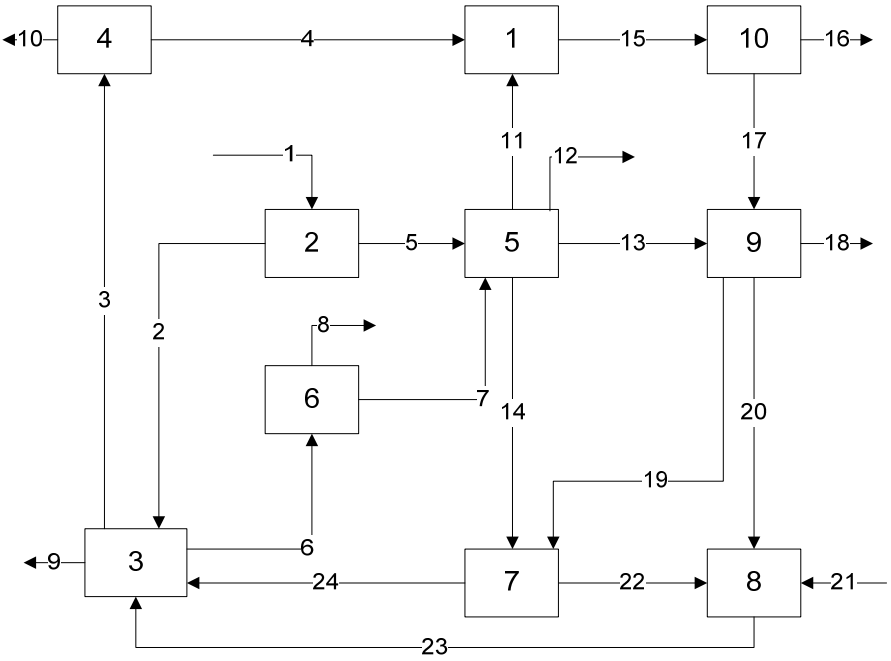

Figura 5.4: Sistema de medição de vapor .

A matriz de incidência é dada por:

$$A = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \quad (5.15)$$

Tabela 5.4: Dados para o Estudo de Caso 2.

N°	x	σ	N°	x	σ	N°	x	σ
1	296.6	35.18	10	1.9	0.001	19	35.7	0.50
2	50.6	1.02	11	84.7	2.86	20	5.7	0.013
3	117.2	5.49	12	70.2	1.97	21	264.6	28.00
4	115.3	5.31	13	132.0	6.96	22	37.1	0.55
5	246.0	24.20	14	202.1	16.33	23	307.4	37.79
6	244.8	23.97	15	200.0	16.00	24	8.7	0.03
7	243.0	23.61	16	188.0	14.13	25	192.0	14.74
8	1.8	0.001	17	12.0	0.057			
9	4.7	0.008	18	102.6	4.21			

*Onde x corresponde ao valor verdadeiro do estado e σ ao desvio padrão de 2,5% utilizado no exemplo.

Este caso é amplamente utilizado na literatura e pode ser encontrado em: Serth e Henan (1986), Narashiman e Mah (1987), Yang et al. (1995), Rollins et al. (1996), Zhang et al. (2000), Soderstrom et al. (2000), Narashiman e Jordache (2000), Romagnoli e Sanchéz (2000), Arora e Biegler (2001), Congli et al. (2006), Mei et al. (2006), Narashiman e Shah (2008), entre outros.

O trabalho utilizado como referência foi publicado por Narashiman e Mah (1987) possibilitando a comparação com Rollins et al. (1996). Assim, para todas as simulações o número de variáveis com erros grosseiros foi escolhida fixada em 3, 5 ou 7. O tamanho do erro grosseiro foi variado entre 5-25 desvios padrões (ou ainda 12,5% - 62,5% do valor real das variáveis), escolhido aleatoriamente com igual probabilidade. A localização dos erros grosseiros também foi escolhida de maneira aleatória, assim como o sinal dos erros. Como em Rollins et al. (1996), a matriz de variância-covariância é diagonal.

Capítulo 6

Resultados e Discussão

Neste capítulo são apresentados os resultados da aplicação da metodologia proposta no Capítulo 5, divididos em 4 partes. As três primeiras partes referem-se à utilização do estudo de caso da Rede de trocadores de calor com reciclo (Rosenberg, 1986). Na quarta parte são apresentados casos da literatura visando validar o novo método – IMT robusto.

Na primeira parte são avaliadas as influências da topologia do processo e do pré-tratamento dos dados sobre a RD. Na segunda parte são aplicadas as técnicas de DEG apresentadas. Num primeiro momento avalia-se a influência da topologia no desempenho da detecção e estimação dos estados. Após são obtidas as curvas de Poder de Detecção Global (OP) em função do tamanho do erro grosseiro e em função do intervalo de confiança dos testes estatísticos. Esta etapa é exploratória e visa preparar para a sintonia das estratégias de maneira que possam ser comparadas de maneira justa. A seguir, as diferentes estratégias são comparadas entre si e a reconciliação robusta.

Na quarta e última etapa é realizada a validação do novo método. Nesta, primeiro é apresentado o desenvolvimento (os testes iniciais, a escolha das premissas e sintonia) e por fim é realizada a comparação com os resultados já obtidos e com os trabalhos existentes na literatura.

6.1 Resultados e Discussão para Reconciliação de Dados

6.1.1 Influência da topologia

Quando existem variáveis não medidas, o conjunto de dados perde tanto redundância quanto observabilidade. Como conseqüência, a reconciliação de dados fornece precisão menor que a desejada. O mesmo acontece quando são utilizadas as estratégias eliminatórias de detecção. Por isto, é realizado um estudo exploratório e combinatório visando entender o quanto dos resultados obtidos para as estratégias de detecção de erros tem explicação na topologia do processo.

Para avaliar a topologia frente à reconciliação de dados é utilizado o conceito de *ajustabilidade*. Este serve para quantificar o impacto de uma variável não medida. Como já dito anteriormente, a ajustabilidade é um coeficiente que relaciona o desvio padrão entre os estados estimados com o desvio padrão das medições e, se menor que um valor crítico (0.1, por exemplo), indica que a variável pode ser considerada não redundante na prática. Isto acontece porque o ajuste realizado na variável medida é insignificante.

Os valores para todas as variáveis medidas (caso 0) já foi mostrado no capitulo anterior. São avaliados neste item os resultados referentes à retirada de uma variável por vez (casos 1 ao 7), à retirada de uma das combinações entre 2 variáveis (casos 8 à 28), ou ainda à retirada de uma das combinações de 3 variáveis (casos 29 à 63). Os resultados obtidos para ajustabilidade (*Aj*), para os casos 0 ao 7 são apresentados na Tabela **6.1**. Nesta são apresentados dois resultados: *Aj* e a redução percentual em relação ao valor do caso 0. Na Figura 6.1 é apresentado o efeito da retirada de cada uma das variáveis na ajustabilidade das variáveis restantes. Neste, os eixos correspondem às variáveis retiradas e cada série corresponde à ajustabilidade.

Tabela 6.1: Ajustabilidade para os casos 0 ao 7.

	N° do Caso														
	0	**1**		**2**		**3**		**4**		**5**		**6**		**7**	
i	**$(Aj)_i$**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**	**$(Aj)_i$**	**%**
1	0.57	-	-	0.56	1.4	0.56	1.4	0.55	3.5	0.54	5.3	0.50	11.6	0.23	59.0
2	0.84	0.83	1.4	-	-	0.81	3.6	0.50	40.6	0.82	2.4	0.74	12.4	0.83	1.4
3	0.84	0.83	1.4	0.81	3.6	-	-	0.50	40.6	0.82	2.4	0.74	12.4	0.83	1.4
4	0.15	0.15	3.5	0.09	40.6	0.09	40.6	-	-	0.14	5.8	0.11	30.6	0.15	3.5
5	0.77	0.73	5.3	0.75	2.4	0.75	2.4	0.73	5.8	-	-	0.42	45.5	0.73	5.3
6	0.26	0.23	11.6	0.23	12.4	0.23	12.4	0.18	30.6	0.14	45.5	-	-	0.23	11.6
7	0.57	0.23	59.0	0.56	1.4	0.56	1.4	0.55	3.5	0.54	5.3	0.50	11.6	-	-

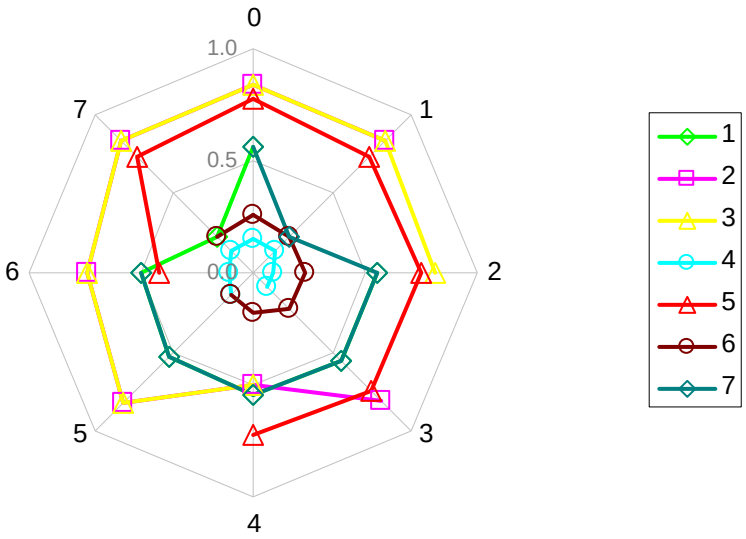

Figura 6.1: Efeito da retirada de uma variável por vez na ajustabilidade das variáveis restantes.

Pode-se notar que as variáveis 1 e 7 apresentam o mesmo comportamento – a retirada de uma representa a redução da ajustabilidade da outra e o valor numérico é exatamente igual. As correntes 5 e 6 apresentam comportamento equivalente, apesar da variável 6 ser bem menos redundante que a 5. Já as variáveis 2 e 3 apresentam o mesmo comportamento de redução de ajustabilidade quando é retirada da variável 4. E esta última só não é influenciada pelas correntes 1 e 7 e, na maioria dos casos, pode ser considerada uma variável não redundante na prática. Na média, a ajustabilidade foi reduzida em 15% do valor obtido no caso 0.

Os resultados obtidos para os casos 8-18 são apresentados na Tabelas 6.2 e para os casos 19-28 na Tabela 6.3. Estes são os casos onde 2 variáveis são consideradas não medidas. As células em cinza têm por finalidade referenciar quais as variáveis não medidas de cada caso. Na Tabela 6.2 também é apresentado o caso 0 para fins comparativos. Analisando os resultados, verifica-se que em 9 casos existem variáveis que se tornaram não redundantes na prática (Aj = 0). Na média, ocorreu uma redução de aproximadamente 30% na ajustabilidade, com a retirada de 2 variáveis.

Tabela 6.2: Ajustabilidade para os casos 8 à 18

	Casos											
i	**0**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	0.57	-	-	-	-	-	-	0.55	0.55	0.52	0.50	0.21
2	0.84	-	0.79	0.50	0.80	0.74	0.74	-	-	-	-	-
3	0.84	0.79	-	0.50	0.80	0.74	0.74	-	0	0.78	0.64	0.79
4	0.15	0.09	0.09	-	0.13	0.11	0.11	0	-	0.09	0.07	0.09
5	0.77	0.71	0.71	0.67	-	0.42	0.42	0.73	0.73	-	0.29	0.71
6	0.26	0.21	0.21	0.17	0.13	-	0	0.18	0.18	0.09	-	0.21
7	0.57	0.21	0.21	0.17	0.13	0	-	0.55	0.55	0.52	0.50	-

Tabela 6.3: Ajustabilidade para os casos 19 à 28

	Casos									
i	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
1	0.55	0.50	0.50	0.21	0.50	0.50	0.17	0.50	0.13	0
2	0	0.78	0.64	0.79	0.50	0.50	0.50	0.50	0.80	0.74
3	-	-	-	-	0.50	0.50	0.50	0.50	0.80	0.74
4	-	0.09	0.07	0.09	-	-	-	0	0.13	0.11
5	0.73	-	0.29	0.71	-	0	0.67	-	-	0.42
6	0.18	0.09	-	0.21	0	-	0.17	-	0.13	-
7	0.55	0.52	0.50	-	0.50	0.50	-	0.50	-	-

Os resultados para os casos 29 à 63, onde são retiradas 3 variáveis por vez, são apresentados em forma de *redução média de ajustabilidade* e se encontram na Tabela 6.4. A *redução média de ajustabilidade* é dada pela média da redução de ajustabilidade para todas as variáveis em cada caso. Para facilitar o entendimento são apresentadas, nas colunas ***u,*** as correntes retiradas em cada caso. Salienta-se que quando o resultado é 100%, existe *perda de observabilidade teórica*. Isto acontece quando o posto da matriz das restrições (*A*) é menor do que o número de variáveis retiradas do sistema, fazendo com que a solução não seja única. Portanto não é possível estimar as variáveis restantes. Existem 3 casos (43, 44, 60) em que o sistema torna-se não observável teoricamente e não pode ser estimado.

Tabela 6.4: Redução média da Ajustabilidade para os casos 29 à 63

caso	u			Aj. %	caso	u			Aj. %	caso	u			Aj. %
29	1	2	3	55	**41**	1	5	6	70	**53**	2	6	7	60
30	1	2	4	55	**42**	1	5	7	70	**54**	3	4	5	56
31	1	2	5	52	**43**	1	6	7	100	**55**	3	4	6	56
32	1	2	6	60	**44**	2	3	4	100	**56**	3	4	7	55
33	1	2	7	60	**45**	2	3	5	56	**57**	3	5	6	56
34	1	3	4	55	**46**	2	3	6	56	**58**	3	5	7	52
35	1	3	5	52	**47**	2	3	7	55	**59**	3	6	7	60
36	1	3	6	60	**48**	2	4	5	56	**60**	4	5	6	100
37	1	3	7	60	**49**	2	4	6	56	**61**	4	5	7	70
38	1	4	5	70	**50**	2	4	7	55	**62**	4	6	7	70
39	1	4	6	70	**51**	2	5	6	56	**63**	5	6	7	70
40	1	4	7	70	**52**	2	5	7	52					

Na Tabela 6.4 observa-se que a redução média na ajustabilidade é de 60%. Isto quês dizer que, quando existirem 3 erros grosseiros neste sistema, o ajuste realizado na reconciliação é em média 60% menor. Conclui-se, então, que este comportamento deve ser esperado nos resultados. À medida que mais erros grosseiros são introduzidos no sistema, a perda de redundância leva a um ajuste realizado pela reconciliação de dados 60% menor.

Este comportamento deve aparecer mais fortemente nas estratégias eliminatórias, visto que as variáveis são tratadas como não medidas. Esta é uma desvantagem apresentada por este tipo de estratégia, e que só pode ser superada com a adição de mais medidores. Como uma opção para diminuir o impacto desta desvantagem é a utilização de pré-tratamento de dados. Este é tratado no próximo item e tem como objetivo diminuir a variabilidade dos dados e mais adiante terá o papel de salientar os erros grosseiros, atenuando o ruído de alta freqüência.

6.1.2 Influência do pré-tratamento de dados

Para verificar a influência da utilização de técnicas de pré-tratamento de dados na reconciliação de dados, foram aplicados os 3 filtros diferentes (FIR, IIR, SVG) nos dados brutos e após foi aplicada a RD. Com isto pretende-se comparar a redução na variância, e na TER, para diferentes seqüências de processamento dos dados. Nos resultados apresentados, quando aparece a palavra *somente*, refere-se somente à etapa. Os outros casos referem-se à variável bruta ("Dados"), ou à variável média ("y*m*").

A redução de variância obtida a partir da aplicação dos filtros pode ser vista na Figura 6.2. Nesta, a variável chamada de y_m equivale à média dos 10.000 pontos para N = 5. A janela para a média é escolhida com base na mesma premissa aplicada à sintonia dos filtros, ou seja, que a redução na variância seja de 80%. Pode-se verificar que os 3 filtros mostram uma redução na variância em torno de 80%. Na Tabela 6.5 são apresentados os valores obtidos em função da Redução Total do Erro (TER) e na Tabela 6.6 são apresentados os resultados obtidos para a variância (σ^2). A diferença entre as Tabelas 6.5 e 6.6 é que, no caso do TER, os resultados referem-se à redução na variância de *todo* o conjunto de dados. Já para a variância

é apresentado o resultado obtido para a corrente 1. Os resultados para as outras correntes seguem o mesmo comportamento já que os filtros são lineares

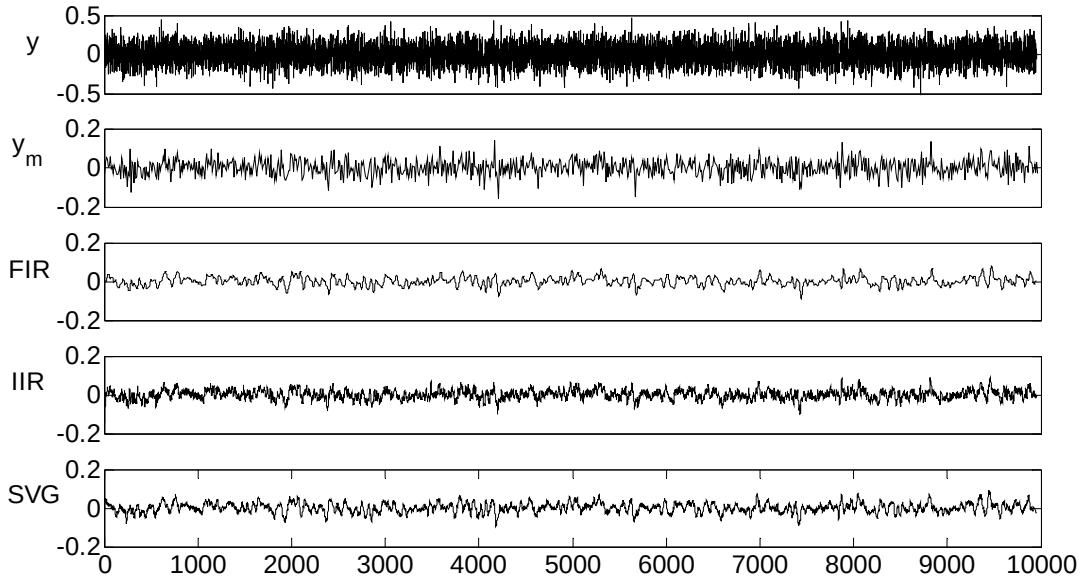

Figura 6.2: Redução do ruído de medição devido à filtragem dos dados brutos.

Tabela 6.5: TER obtido para as diferentes etapas

ETAPAS	TER		
Reconciliação de y	0.38		
Reconciliação da Média de y	0.24		
FILTROS	FIR	IIR	SVG
Somente Filtragem de y	**0.80**	**0.78**	**0.78**
Somente reconciliação de y filtrado	0.22	0.22	0.23
(Filtragem + Reconciliação)	**0.85**	**0.83**	**0.83**
(Filtragem + Média + Reconciliação)	0.22	0.14	0.23
(Média + Reconciliação + Filtragem)	0.32	0.45	0.52
Somente reconciliação da y_m filtrada	0.17	0.18	0.17
(Média + Filtragem + Reconciliação)	**0.85**	**0.79**	**0.83**

Tabela 6.6: Variância obtida para as diferentes etapas

ETAPAS	σ^2		
Reconciliação de y	0.016		
Reconciliação da Média de y	0.00067		
FILTROS	FIR	IIR	SVG
Filtragem de y	0.00021	0.00025	0.00025
Reconciliação de y filtrado	0.00020	0.00019	0.00023
Filtragem da média	0.000036	0.00019	0.000049
(Média + Filtragem + Reconciliação)	0.000023	0.00017	0.000027

Para avaliar a distribuição dos dados filtrados, são apresentadas na Figura 6.3 as gaussianas obtidas a partir dos dados filtrados, brutos e da média. Estas figuras foram obtidas utilizando-se a função *normfit* do *Matlab.* As escalas estão distintas para permitir a visualização do formato das curvas.

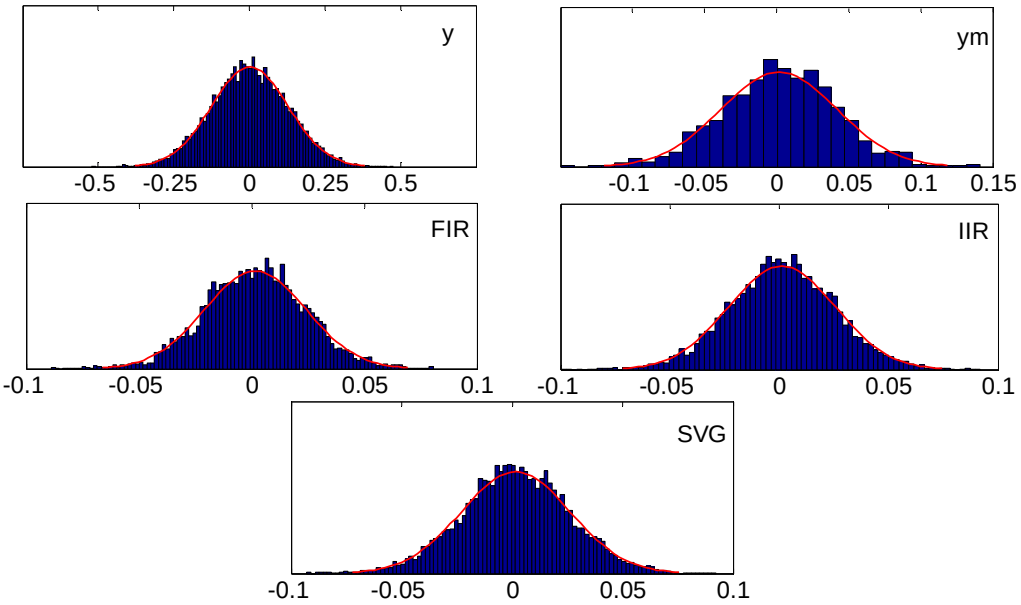

Figura 6.3: Gaussianas geradas a partir dos dados filtrados (FIR, IIR e SVG), brutos (y) e da média (ym).

Como se pode ver, a distribuição continua aproximadamente normal, e somente ocorreu a diminuição da variância dos dados. Este resultado é explicado pelo fato dos filtros realizarem somente operações lineares nos dados. A única das premissas que pode ser questionada é do ruído ser branco, pois este apresenta correlação, mesmo que pequena na dimensão temporal do sinal (o que não influencia na reconciliação de dados estacionária). A partir dos dados filtrados, aplicou-se então a reconciliação de dados. Os resultados para a variável 1 são apresentados na Figura 6.4. A variável reconciliada é chamada de x_{est}.

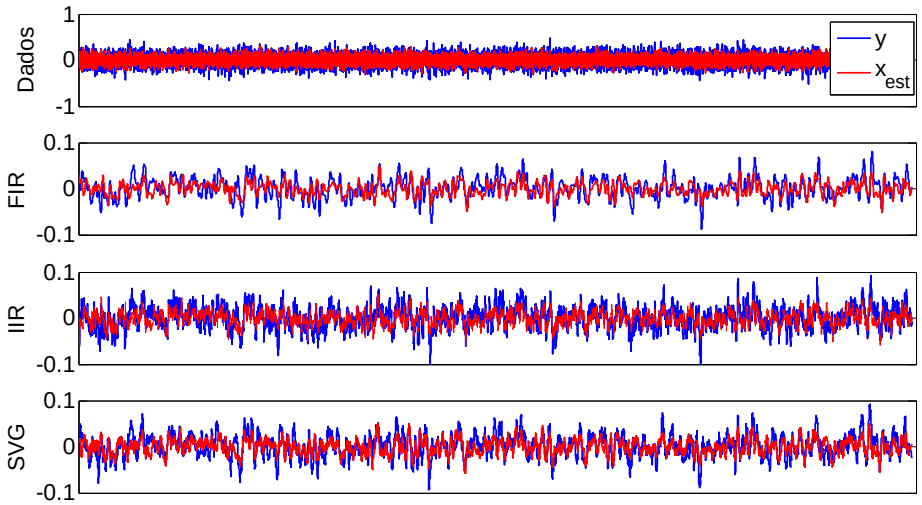

Figura 6.4: Comparação entre o ruído de medição presente nos dados filtrados e brutos (y) com o ruído após a reconciliação (X_{est}).

Como esperado, a aplicação da reconciliação gera uma segunda diminuição da variância dos dados. As gaussianas geradas após a seqüência (filtragem – reconciliação) são

apresentadas nas Figuras 6.5 e 6.6. A primeira figura tem o objetivo de mostrar o formato e na segunda mostra-se a escala. Na Figura 6.5 observa-se que os dados reconciliados apresentam a forma desejada. Já na Figura 6.6 tem-se a noção da redução acentuada da variância obtida a partir da filtragem seguida da reconciliação de dados. Todos os filtros apresentam resultados equivalentes para a reconciliação estacionária.

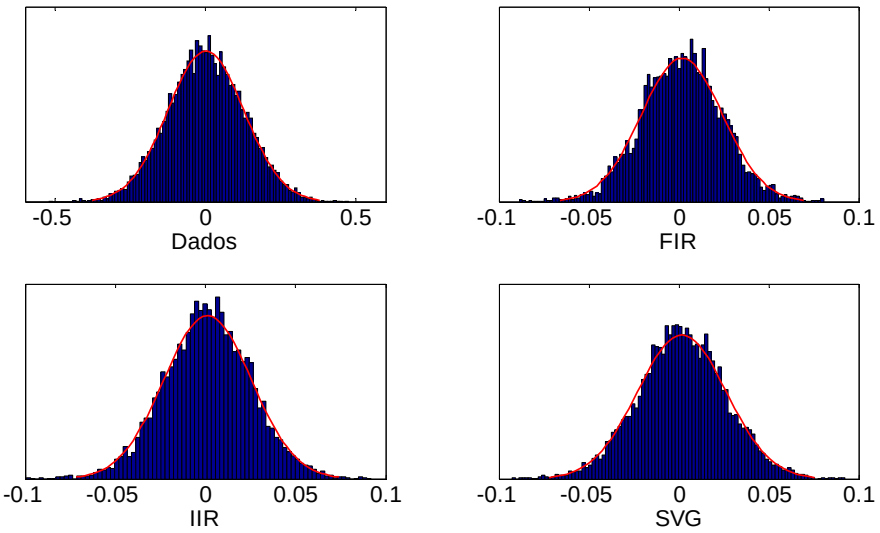

Figura 6.5: Comparação entre as gaussianas obtidas a partir dos dados brutos (Dados) e dos dados filtrados e reconciliados (FIR, IIR e SVG).

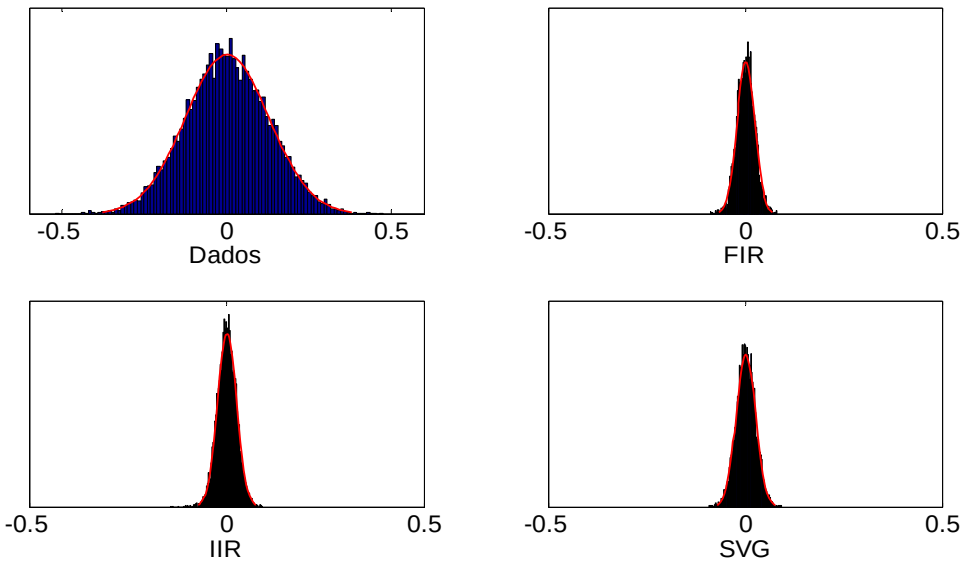

Figura 6.6: Gaussianas da Figura 6.5 na mesma escala de visualização

Foi também avaliada a utilização da filtragem dos dados a partir de um conjunto médio de dados de processo. Esta é uma situação comum em dados industriais, e se dá quando são fornecidos dados médios do processo e pede-se que seja feita a reconciliação. A idéia é verificar se é possível aplicar tal procedimento e se é vantajoso. Para isto, foram obtidas as

médias dos mesmos dados utilizados na etapa anterior (com redução no tamanho da amostra) e estes foram filtrados utilizando-se os mesmos filtros já apresentados. Para demonstrar que os filtros retiram mais componentes de alta freqüência (ruído branco aleatório), são apresentados os espectros de potência para filtragem da média dos dados e as gaussianas (em escalas diferentes) obtidas nas Figuras 6.7 e 6.8.

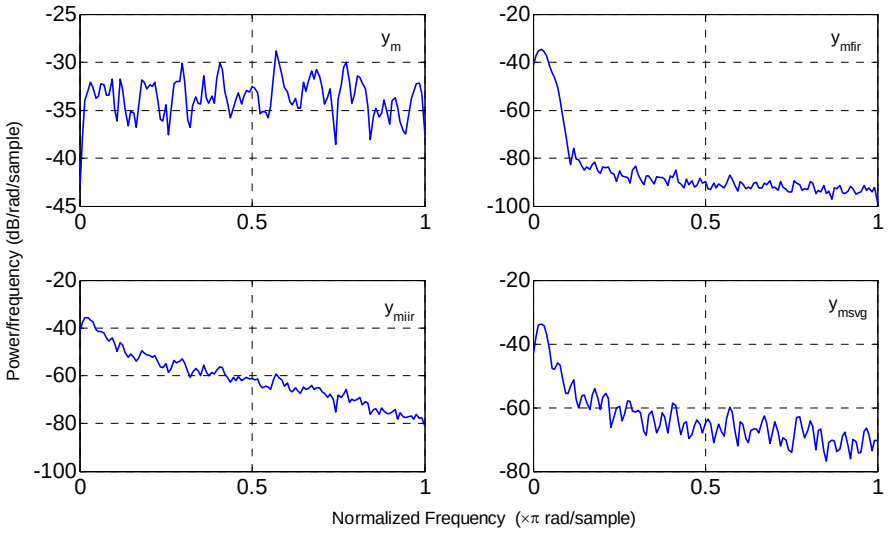

Figura 6.7: Espectro de potência da 2° etapa (Filtragem da média).

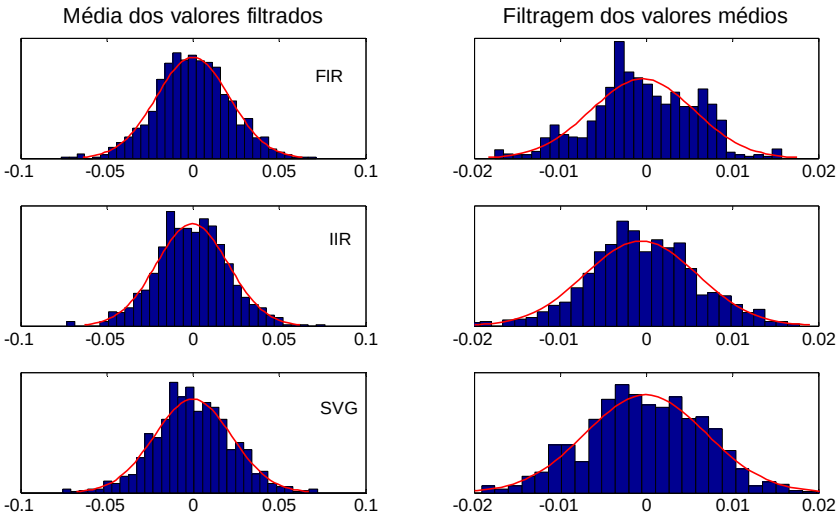

Figura 6.8: Comparação entre as gaussianas obtidas a partir da média dos dados filtrados e da filtragem da média dos dados.

Após foi aplicado o procedimento de reconciliação de dados aos dados médios filtrados. Na Figura 6.9 são apresentadas as gaussianas obtidas. Nesta estão as gaussianas obtidas a partir da média dos valores filtrados (1° comparação realizada) e os resultados obtidos nesta segunda comparação.

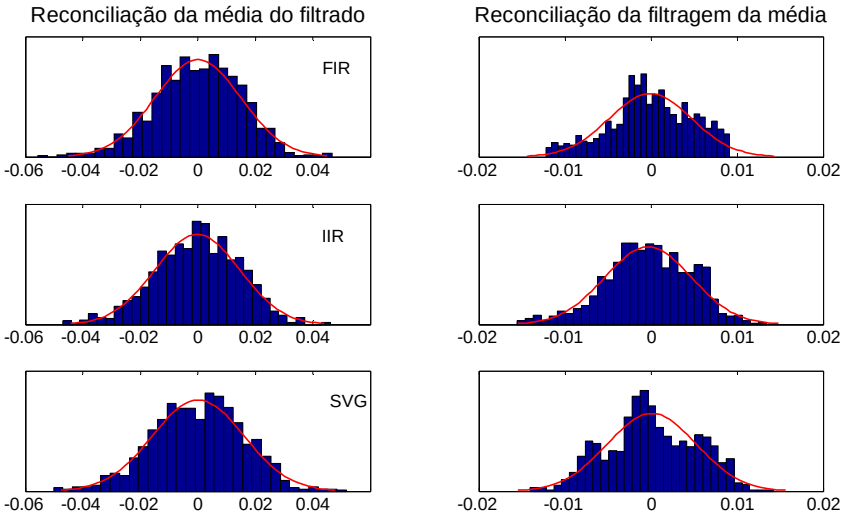

Figura 6.9: Resultados para a reconciliação da média da filtragem e da filtragem da média.

Pode-se ver que a quantidade menor de pontos utilizada pela filtragem da média faz com que a curva não fique tão característica. Apesar disto, as curvas resultantes são centradas em zero e apresentam a variância menor que na primeira comparação, demonstrando que é possível aplicar a técnica em dados médios e que pode-se esperar algum ganho, principalmente nos algoritmos de detecção de erros grosseiros.

Neste item foram apresentados os resultados para filtragem em relação ao procedimento de reconciliação propriamente dito e como conclusão tem-se que o procedimento é vantajoso. Mais adiante estes resultados serão reforçados em relação aos algoritmos de detecção de erros grosseiros. Com isto espera-se ter uma avaliação mais completa sobre a utilização deste tipo de técnica para auxiliar a detecção de erros grosseiros.

6.2 Resultados e Discussão para Detecção de Erros Grosseiros

Para a avaliação das estratégias de detecção são avaliados três tópicos que influenciam nos resultados das estratégias de detecção: topologia, nível de confiança e tamanho do erro grosseiro e a utilização do pré-tratamento dos dados. Este item será então dividido de maneira que, na primeira parte estuda-se a influência da topologia do processo. Neste será demonstrado o efeito “smearing” e a influência sobre a detecção. Já na segunda etapa serão demonstradas as curva de poder de detecção para os diferentes métodos, visando avaliar a influência do nível de confiança para os testes estatísticos e o tamanho do erro grosseiro.

A partir destes estudos são definidos os critérios para a comparação entre as estratégias de detecção. São então apresentados os resultados comparativos entre os diferentes métodos de detecção. Após são comparados os resultados com os obtidos utilizando a

reconciliação robusta de dados. Na terceira etapa será demonstrado o efeito do pré-tratamento de dados sobre os índices de desempenho da detecção.

6.2.1 Influência da Topologia

Este item tem como objetivo, além de avaliar o impacto da topologia, validar a maneira com que as simulações para comparação entre as estratégias serão realizadas mais adiante. Como já dito anteriormente, o *espalhamento* do erro grosseiro depende de 3 fatores: a topologia do processo em si, o tamanho do erro grosseiro e a relação entre a precisão dos medidores. No item 6.1.1 já se pode ter uma idéia do comportamento da detecção frente à topologia, mas não são apresentados graficamente os resultados relativos à adição de erros grosseiros no sistema.

Para verificar a influência dos parâmetros associados ao espalhamento, foram realizados 3 conjuntos de simulações. De modo a facilitar a interpretação dos resultados, não foi adicionado ruído aleatório nas medições. No primeiro conjunto, influência da topologia é avaliada. Para este fim, foi adicionado um erro grosseiro, com tamanho de 10σ, em uma das variáveis. Os resultados obtidos estão na Figura 6.10.

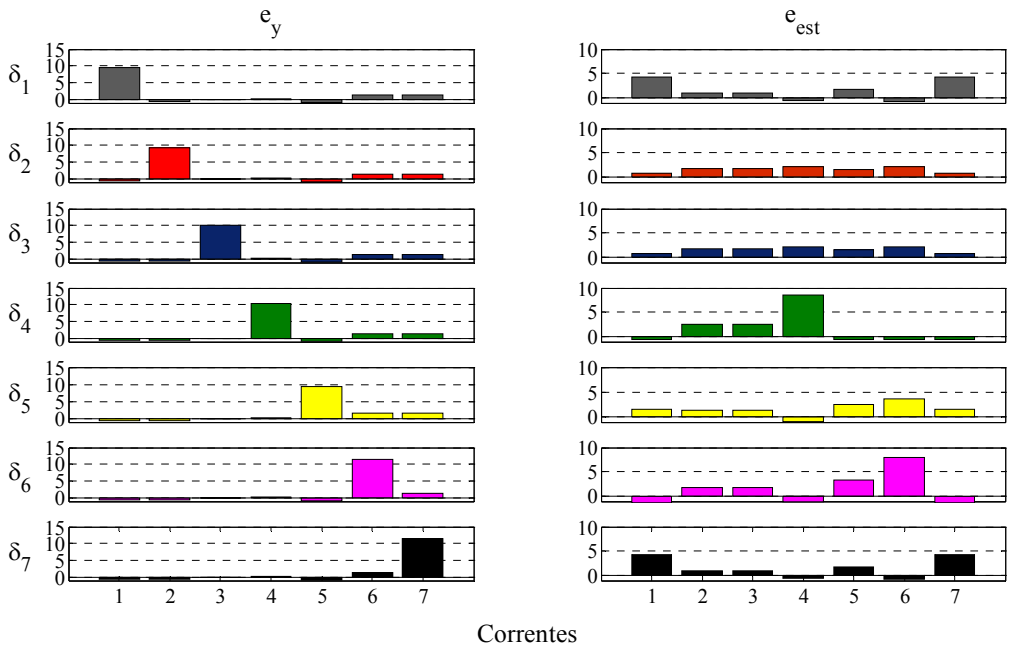

Figura 6.10: Efeito "smearing" em função da topologia do processo.

Na Figura 6.10 e nas duas subseqüentes os gráficos são apresentados em função do erro normalizado da medição, e_y, e do erro normalizado da reconciliação, e_{est}, dados por:

$$e_y = \frac{(y - x)}{\sigma} \qquad e_{est} = \frac{(x_{est} - x)}{\sigma} \tag{6.1}$$

Como se pode ver na Figura 6.10, os resultados refletem os já obtidos na etapa de avaliação da topologia sobre a reconciliação de dados. Os erros são espalhados para os medidores mais próximos e em função da precisão destes. Por exemplo, as variáveis 2 e 3 sempre recebem algum erro, pois o desvio padrão destas é maior do que das outras. Já durante a reconciliação com a variável 4 contendo erro grosseiro o erro é pouco espalhado para o conjunto de dados, quando comparado com o que acontece quando as variáveis 2 e 3 contêm o erro, visto que esta é considerada não redundante na prática.

O segundo conjunto de simulações refere-se à avaliação da influência do tamanho do erro grosseiro adicionado. Portanto, foram adicionados 7 erros grosseiros, de diferentes magnitudes, na variável 1. Na Figura 6.11 são apresentados os resultados.

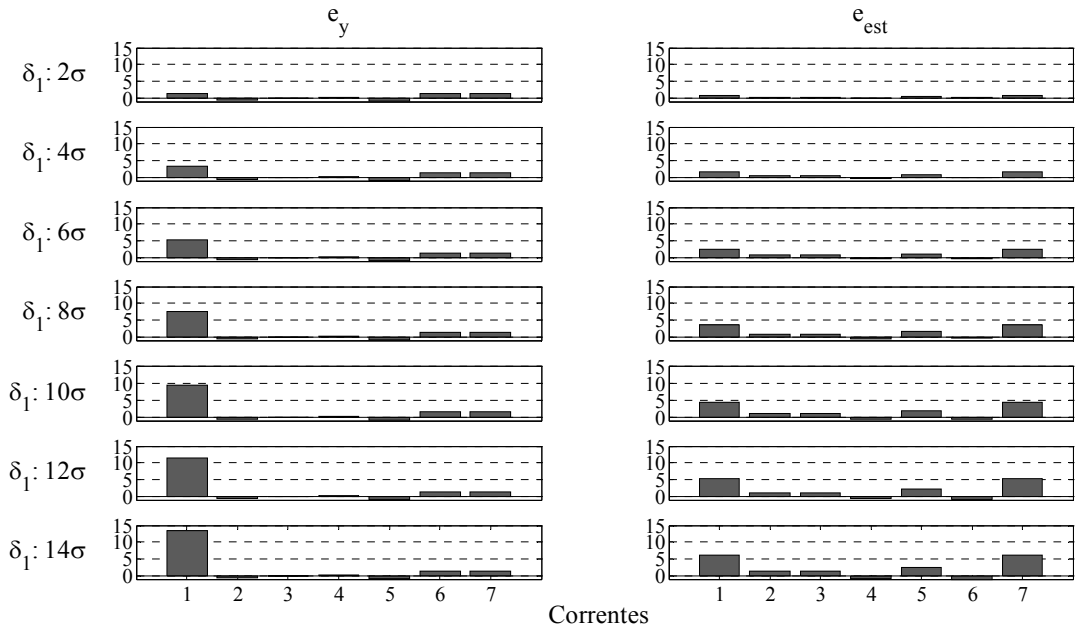

Figura 6.11: Efeito "smearing" em função do tamanho do EG.

Estes resultados justificam a maneira com que os testes de desempenho para as estratégias de detecção são realizados. E também justificam a concepção do algoritmo IMT robusto, visto que o "smearing" afeta fortemente as variáveis livres de erros, a ponto de gerar grandes desvios, introduzindo erros grosseiros ao processo. Este desvio é também chamado de *bias induzido* por Bagajewicz (1997, 2000, 2003).

Já o terceiro conjunto de simulações foi realizado com o objetivo de avaliar a influência da confiança do medidor (que é representada pelo seu desvio padrão, e conseqüentemente a sua variância). Os resultados são mostrados na Figura 6.12. Infelizmente este é um dos parâmetros que não podem ser explorados nesta dissertação, visto que está mais relacionado com a área de projeto de malhas de medição. Os outros dois parâmetros são explorados, indiretamente, de modo que as simulações apresentem igual probabilidade de todas as situações ocorrerem, e assim, os dados de desempenho obtidos para as estratégias de detecção são significativos.

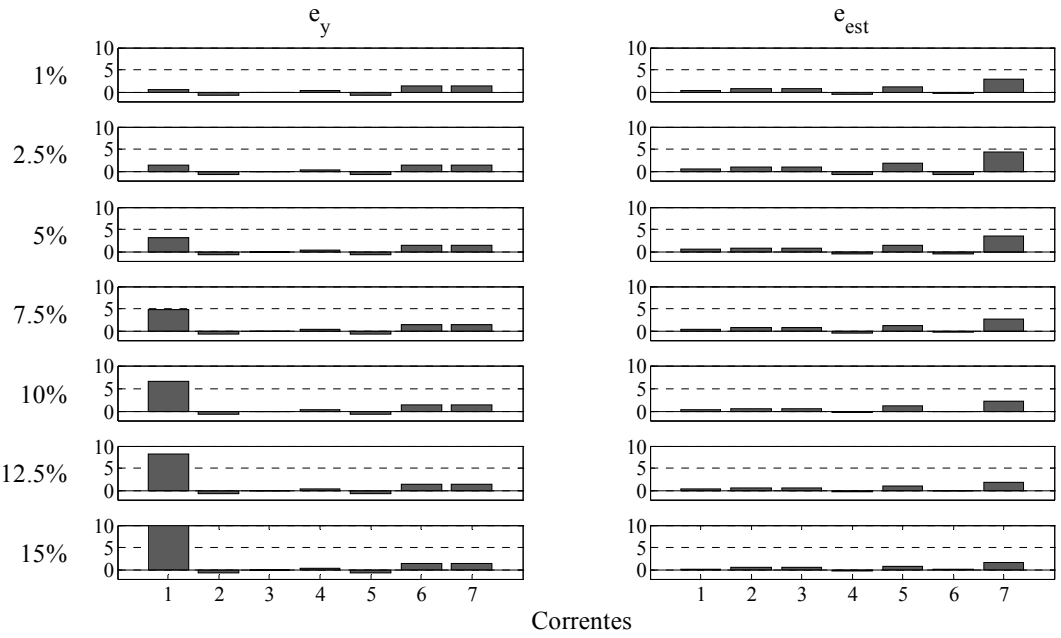

Figura 6.12: Efeito "smearing" em função da precisão do sensor.

Para avaliar o efeito da topologia na detecção de erros grosseiros utiliza-se o conceito de *detectabilidade (Dt)*, que é uma medida de *observabilidade*. Quanto maior o coeficiente, mais facilmente erros grosseiros menores serão detectados nesta variável. Como já demonstrado no Capítulo 2, o cálculo da detectabilidade é equivalente ao da ajustabilidade, sendo que o primeiro é a relação entre as variâncias entre os valores estimados pela reconciliação e a medição e o segundo entre os desvios padrões. Equivalente ao demonstrado no item 6.1.1, os valores referentes à retirada de uma variável por vez (casos 1 à 7), uma das combinações entre 2 variáveis (casos 8 ao 28), ou ainda, uma das combinações de 3 variáveis (casos 29 ao 63) são avaliados neste item. Os resultados para detectabilidade (Dt) e a redução percentual em relação ao caso 0, para os casos 1 ao 7, são apresentados na Tabela 6.7.

O tratamento da variável como uma variável não medida é um procedimento comum nos algoritmos de detecção de erros grosseiros, mas, como pode ser visto nos resultados, tem como desvantagem a perda da observabilidade. Na Figura 6.13, é apresentada a redução percentual na observabilidade com a retirada das variáveis, que reflete todas as relações apresentadas.

Tabela 6.7: Detectabilidade para os casos 0 ao 7.

	Caso														
	0	**1**		**2**		**3**		**4**		**5**		**6**		**7**	
i	**Dt**	**Dt**	**%**	**Dt**	**%**	**Dt**	**%**	**Dt**	**%**	**Dt**	**%**	**Dt**	**%**	**Dt**	**%**
1	0.90	-	-	0.90	0.4	0.90	0.4	0.89	1.1	0.89	1.7	0.87	3.8	0.64	28.9
2	0.99	0.99	0.2	-	-	0.98	0.5	0.87	12.3	0.98	0.3	0.96	2.3	0.99	0.2
3	0.99	0.99	0.2	0.98	0.5	-	-	0.87	12.3	0.98	0.3	0.96	2.3	0.99	0.2
4	0.53	0.52	1.7	0.41	21.6	0.41	21.6	-	-	0.52	2.7	0.45	15.7	0.52	1.7
5	0.97	0.96	1.1	0.97	0.5	0.97	0.5	0.96	1.2	-	-	0.82	16.3	0.96	1.1
6	0.67	0.64	5.2	0.64	5.5	0.64	5.5	0.57	14.8	0.52	23.7	-	-	0.64	5.2
7	0.90	0.64	28.9	0.90	0.4	0.90	0.4	0.89	1.1	0.89	1.7	0.87	3.8	-	-

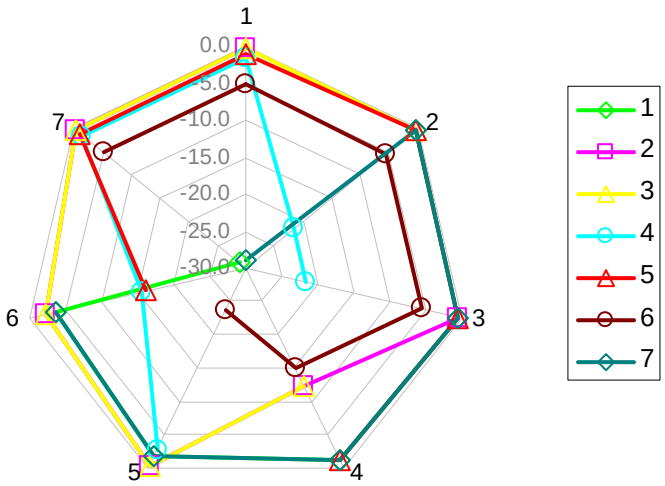

Figura 6.13: Redução percentual na detectabilidade em função de uma variável não medida.

Na Figura 6.13, nota-se que as correntes 1 e 7 apresentam o mesmo comportamento: a retirada de uma gera uma redução na detectabilidade da outra. Além disto, o valor numérico da detectabilidade é exatamente igual. As correntes 5 e 6 apresentam comportamento equivalente, apesar da variável 6 apresentar menos redundância que a 5. Já a detectabilidade das correntes 2 e 3 depende da retirada da corrente 4. Esta última, por sua vez, só não é influenciada pelas correntes 1 e 7 e, na maioria dos casos, pode ser considerada uma variável não redundante na prática. Na média, a detectabilidade foi reduzida em 5%.

Os resultados obtidos para a *redução média na detectabilidade* para os casos 8-28 (retirada de 2 variáveis) são apresentados nas Tabelas 6.8 e 6.9. Observa-se que, na média, ocorreu uma redução de aproximadamente 20% na detectabilidade, com a retirada de 2 variáveis ao mesmo tempo e algumas variáveis passaram a não ser mais observáveis. Salienta-se que quando a redução na detectabilidade for 100%, refere-se à perda de observabilidade teórica, onde o posto da matriz das restrições (A) é menor do que o número de variáveis retirada do sistema. Isto quer dizer que a solução não é única e, portanto, não é possível estimar as variáveis restantes. Um terceiro erro grosseiro não pode ser detectado caso se encontre nesta corrente.

Tabela 6.8: Detectabilidade para os casos 8 à 18

	Casos											
i	**0**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	0.57	-	-	-	-	-	-	1.1	1.1	2.5	3.8	32.5
2	0.84	-	0.9	12.3	0.8	2.3	2.3	-	-	-	-	-
3	0.84	0.9	-	12.3	0.8	2.3	2.3	-	100	1.1	5.4	0.9
4	0.15	22.4	22.4	-	5.8	15.7	15.7	100	-	23.0	29.9	22.4
5	0.77	1.9	1.9	3.2	-	16.3	16.3	1.2	1.2	-	28.1	1.9
6	0.26	9.9	9.9	18.1	26.1	-	100	14.8	14.8	39.6	-	9.9
7	0.57	32.5	32.5	38.6	44.6	100	-	1.1	1.1	2.5	3.8	-

Tabela 6.9: Detectabilidade para os casos 19 à 28

	Casos									
i	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
1	1.1	2.5	3.8	32.5	3.8	3.8	38.6	3.8	44.6	100.0
2	100.0	1.1	5.4	0.9	12.3	12.3	12.3	12.3	0.8	2.3
3	-	-	-	-	12.3	12.3	12.3	12.3	0.8	2.3
4	-	23.0	29.9	22.4	-	-	-	100.0	5.8	15.7
5	1.2	-	28.1	1.9	-	100.0	3.2	-	-	16.3
6	14.8	39.6	-	9.9	100.0	-	18.1	-	26.1	-
7	1.1	2.5	3.8	-	3.8	3.8	-	3.8	-	-

Os resultados para os casos 29 – 63 são apresentados na Tabela 6.10. Existem 3 casos (43, 44, 60) em que o sistema torna-se não observável teoricamente e não pode ser estimado. Pode-se notar que a redução na observabilidade é de aproximadamente 50%. Isto equivale a dizer que, se existir mais algum erro grosseiro, este tem 50% a menos de chance de ser detectado (exceto para os casos 43, 44 e 60, que já são não observáveis), ou ainda, os erros grosseiros devem apresentar maior magnitude para que possam ser detectados.

Tabela 6.10: Redução média da detectabilidade para os casos 29 à 63

caso	**u**			**Dt. %**	**caso**	**u**			**Dt. %**	**caso**	**u**			**Dt. %**
29	1	2	3	40	**41**	1	5	6	56	**53**	2	6	7	41
30	1	2	4	40	**42**	1	5	7	56	**54**	3	4	5	52
31	1	2	5	31	**43**	1	6	7	100	**55**	3	4	6	52
32	1	2	6	41	**44**	2	3	4	100	**56**	3	4	7	40
33	1	2	7	41	**45**	2	3	5	52	**57**	3	5	6	52
34	1	3	4	40	**46**	2	3	6	52	**58**	3	5	7	31
35	1	3	5	31	**47**	2	3	7	40	**59**	3	6	7	41
36	1	3	6	41	**48**	2	4	5	52	**60**	4	5	6	100
37	1	3	7	41	**49**	2	4	6	52	**61**	4	5	7	56
38	1	4	5	56	**50**	2	4	7	40	**62**	4	6	7	56
39	1	4	6	56	**51**	2	5	6	52	**63**	5	6	7	56
40	1	4	7	56	**52**	2	5	7	31					

Com esta análise já se pode ter uma idéia do que é interessante esperar de alguns dos algoritmos de detecção de erros grosseiros. Na literatura (Romagnoli, 2000; Narashiman, 2000; Bagajewicz, 1997 e outros) existem autores que afirmam que as estratégias eliminatórias apresentam desempenho inferior aos compensatórios devido à perda de observabilidade e redundância, o que vem ao encontro com os resultados apresentados neste item. Além disto, verifica-se que existem situações em que é pouco provável que um erro grosseiro seja determinado, ou, se determinado, o conjunto de dados não sofre quase ajuste na reconciliação. Este comportamento será levado em consideração mais adiante, durante a avaliação das estratégias de detecção de modo que será aplicado o conceito de erros equivalentes durante o cálculo das estatísticas de desempenho das estratégias.

6.2.2 Influência de α e δ: Curvas de Poder de detecção e AVTI

Como já explicitado anteriormente, o desempenho dos métodos de detecção é fortemente influenciado pelo efeito "smearing" (que, por sua vez, depende da topologia, da precisão da medição e do tamanho do erro grosseiro) e pelas características do teste de

hipótese aplicado (nível de confiança e tipo de hipótese testada). Dois parâmetros que facilmente podem ser manipulados e avaliados são o tamanho do erro grosseiro e o nível de confiança do teste estatístico. Na prática, ambos são definições do analista e servem perfeitamente para avaliar e prever o comportamento das estratégias, visando à escolha mais adequada.

Para avaliar as etapas de detecção e localização dos erros grosseiros e para se ter uma idéia do comportamento dos algoritmos em função do tamanho do erro grosseiro, foram levantadas as curvas de poder para $\alpha = 0.05$. Foram realizadas simulações Monte Carlo, na presença de um único erro grosseiro, com igual probabilidade de qualquer uma das variáveis conterem o erro. Com base nestas foram calculados os indicadores de desempenho OP e AVTI em função do tamanho do erro grosseiro (δ). Os resultados de OP são apresentados na Figura 6.14 para os algoritmos de detecção de um único erro grosseiros: GT, MT, PCA e GLR.

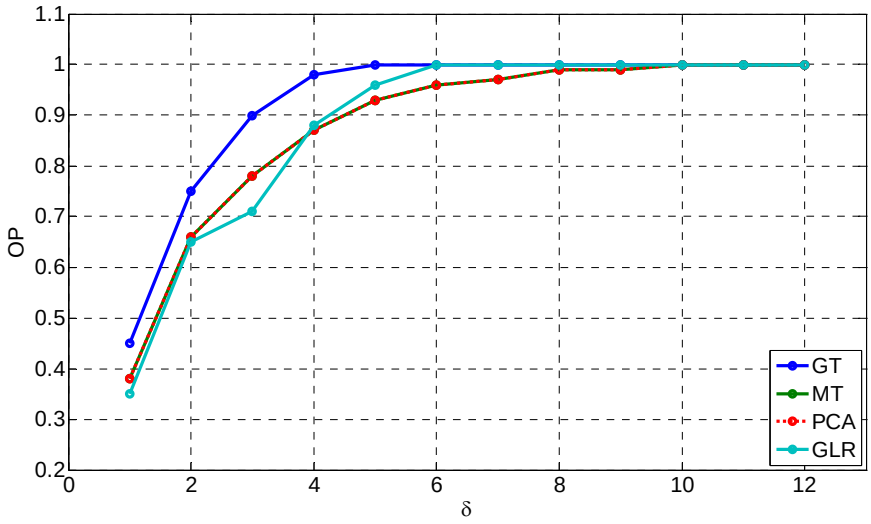

Figura 6.14: Curvas de OP em função do tamanho do EG (δ) dos algoritmos de detecção de um único EG.

Na Figura 6.14 pode-se observar que o algoritmo PCA apresenta um comportamento semelhante ao MT. Narashiman et al. (2000) concluíram que o PCA, quando comparado com o MT, apresenta um poder de detecção teórico menor. Por este motivo, nesta dissertação, em todos os algoritmos que possuem a opção de serem utilizados tanto com o MT quanto com o PCA, se opta pelo MT. Segundo Tong et al. (1995), o PCA poderia apresentar vantagens frente ao MT no caso em que existe correlação entre as variáveis. Entretanto, como uma das premissas desta dissertação é que esta não existe e então o desempenho esperado é o mesmo obtido pelo MT.

Um procedimento similar foi realizado para gerar as curvas das estratégias de detecção de múltiplos erros grosseiros. Para o SEGE, o número máximo de erros grosseiros foi fixado em 2. As curvas de Poder de Detecção Global (OP) obtidas em função do tamanho do EG (δ) são ilustradas na Figura 6.15.

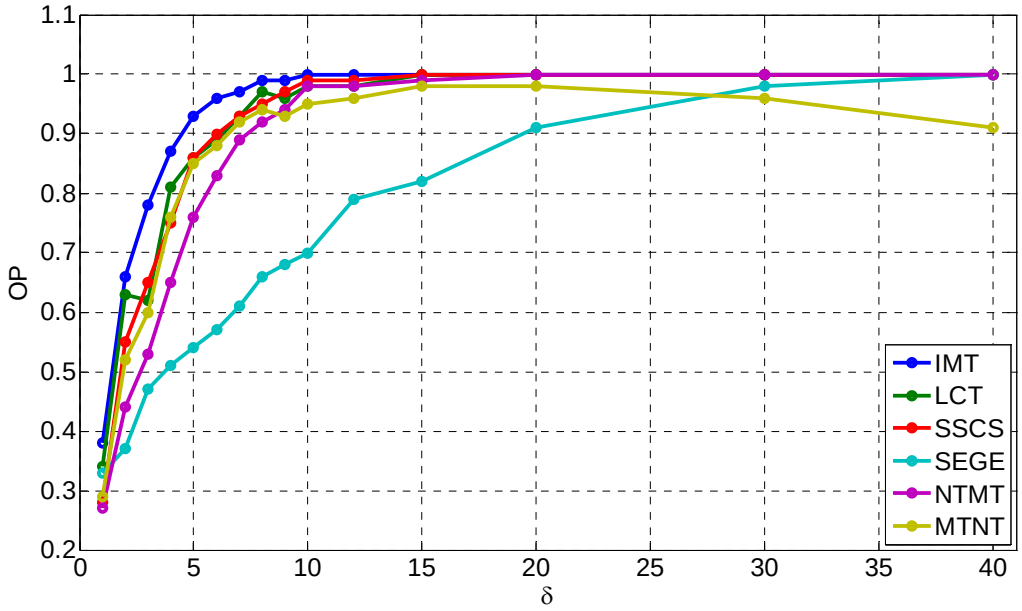

Figura 6.15: Curvas de OP em função de δ das estratégias de detecção de múltiplos EGs.

Com base nestes resultados verifica-se que a maioria dos algoritmos apresenta um padrão de detecção em função do tamanho do erro grosseiro muito parecido. O algoritmo SEGE apresentou o pior desempenho e o MIMT não foi apresentado por possuir o mesmo resultado que o IMT. Infelizmente não existem curvas como estas na literatura de reconciliação de dados para que se possa fazer uma comparação. Este é um procedimento mais comum na literatura de Estatística Avançada e é muito interessante, pois permite que se tenha uma idéia mais clara do comportamento dos testes para a escolha do mais adequado, apesar do trabalho exaustivo de levantamento das curvas.

Entretanto, a avaliação das curvas de poder por si só não garante uma boa análise. Para todos os testes estatísticos pode-se aumentar o poder de detecção até atingir o máximo, mas com o custo de se cometer erros do tipo falso alarme (Tipo I). Portanto, devem ser levados em consideração os valores de AVTI. As curvas para AVTI, assim como a relação OP/AVTI, são apresentadas na Figura 6.16. Curvas do tipo OP/AVTI são comumente utilizadas em detecção clássica aplicada a telecomunicações e detecção de falhas e também são conhecidas como *curvas corr*.

Avaliando os resultados apresentados nas Figuras 6.15 e 6.16 se observa que, apesar do IMT apresentar o maior AVTI, este também apresenta um poder de detecção bem superior aos outros algoritmos testados, fazendo com que a relação entre o OP/AVTI seja a maior de todas. O mesmo acontece com o MTNT e com o NTMT. Para estes dois algoritmos não existem dados na literatura de OP e AVTI provenientes de simulações Monte Carlo. Nos artigos em que foram publicados só aparecem os resultados de uma (no máximo duas) realização e, desta forma, os resultados não podem ser comparados. O que se pode afirmar é que o resultado obtido por estes dois algoritmos é surpreendente. À medida que o tamanho do erro grosseiro cresce, estes algoritmos passam a cometer menos erros do tipo I, no mesmo patamar do IMT. Isto provavelmente se deve à utilização dos testes NT e MT em conjunto.

Já para o LCT e para o SSCS, apesar de um poder de detecção próximo do IMT, a relação OP/AVTI faz com que o desempenho não seja tão satisfatório. Isto também aparece nos dados encontrados em Rollins et al. (1996). Para o SEGE, à medida que o tamanho do erro grosseiro aumenta este comete mais erros e o poder de detecção é muito inferior aos dos outros algoritmos. Isto contradiz o publicado em Sanchez et al. (1999).

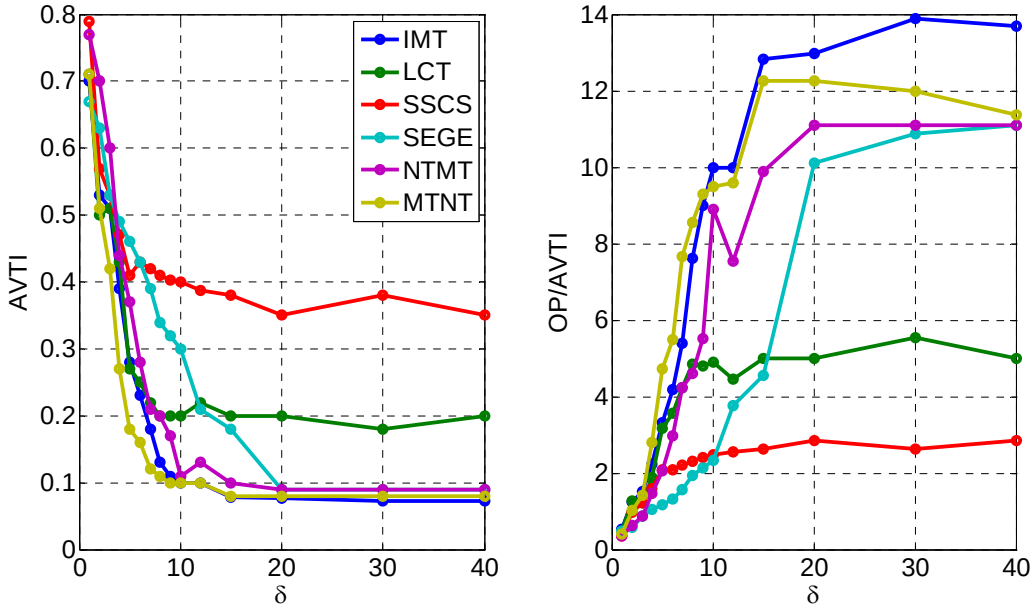

Figura 6.16: Curvas de AVTI em função de δ para as estratégias de detecção de múltiplos EGs.

Além de avaliar os indicadores OP e AVTI, pode-se levar em consideração também a qualidade de estimação das estratégias, expressada pelo indicador de Redução Total do Erro (TER). Na Figura 6.17 são apresentadas as curvas de TER em função de δ. Nesta figura verifica-se que os algoritmos diferem um pouco na qualidade final da estimação. Os que se destacam são o NTMT, o MTNT e o IMT. O SSCS e o LCT apresentam desempenho inferior já que cometem mais erros do tipo I e, conseqüentemente, a etapa de estimação perde a propriedade de ser "unbiased".

Outra explicação para as diferenças em OP e AVTI entre os algoritmos seria o ajuste do nível de confiança. Para um mesmo nível de confiança, nem todas as estratégias apresentam um mesmo patamar de AVTI, fazendo que esta comparação não seja necessariamente justa.

Para verificar a influência do nível de confiança, foram simulados casos variando α e o tamanho do erro grosseiro (δ), para todas as estratégias de detecção. Os resultados para OP são apresentados na Figura 6.18.

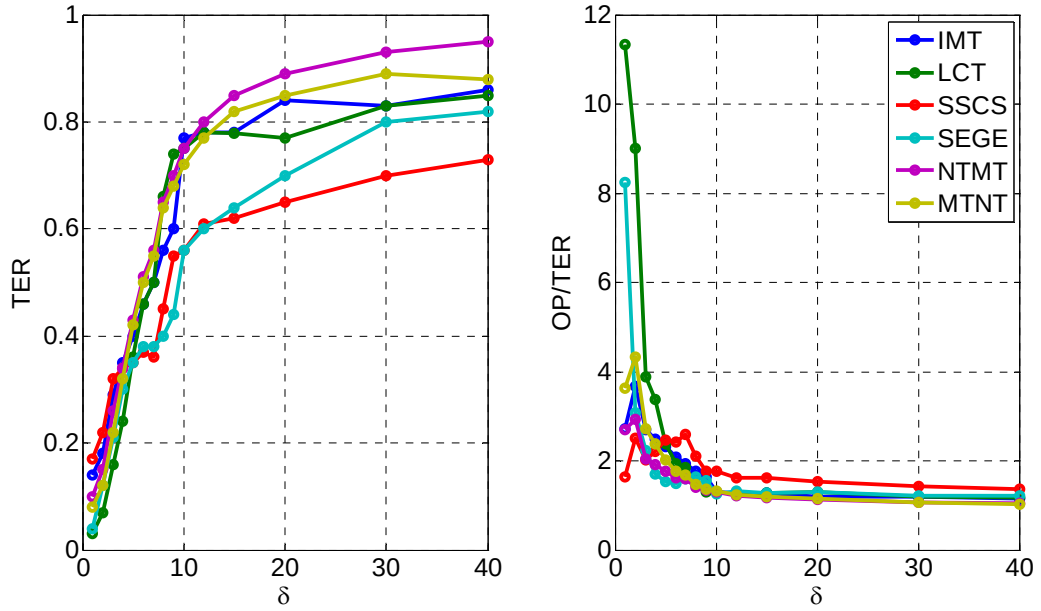

Figura 6.17: Curvas de TER em função de δ para as estratégias de detecção de múltiplos EGs.

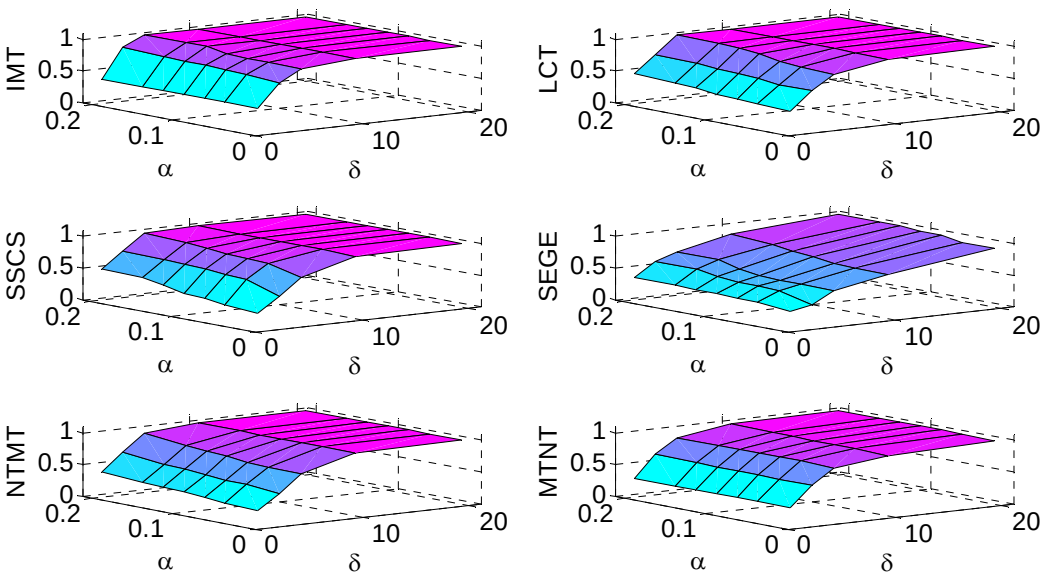

Figura 6.18: Curvas de OP em função de δ e α para as diferentes estratégias de DEG.

Na Figura 6.18 observa-se que as curvas são muito parecidas, e que a diminuição no nível de confiança (aumento de α) faz com que OP aumente. Isto vem ao encontro com o que já foi dito na introdução do Capítulo 4. Existem duas maneiras de se aumentar o poder de detecção de um teste estatístico: aumentando o valor de α ou reduzindo a variância do conjunto de dados. Nota-se que o SEGE apresenta um poder de detecção levemente inferior aos outros, mesmo variando α. Além disto, pode-se observar que este é o método mais sensível à variação deste parâmetro. Já os outros algoritmos apresentam curvas similares,

atingindo valores próximos a 1, mais rapidamente quando comparado ao SEGE. Como já dito anteriormente, nestes casos, a avaliação da curva de poder por si só não é suficiente. Então, nas próximas figuras são apresentados os resultados para AVTI e para a relação OP/AVTI, para todos os algoritmos. As Figuras 6.19, 6.20, 6.21, 6.22, 6.23 e 6.24 referem-se aos algoritmos IMT, MTNT, NTMT, LCT, SEGE e SSCS, respectivamente.

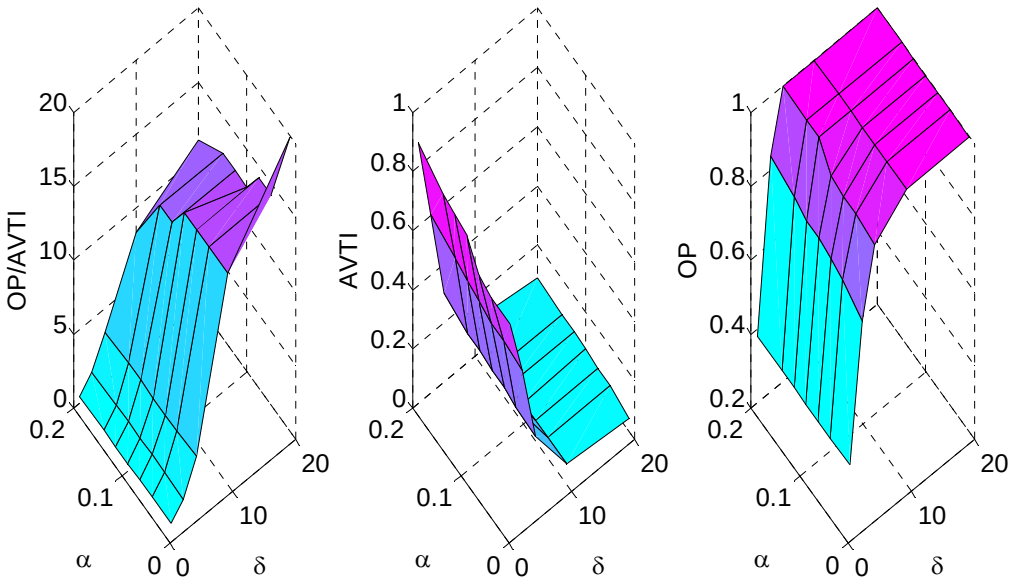

Figura 6.19: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para IMT

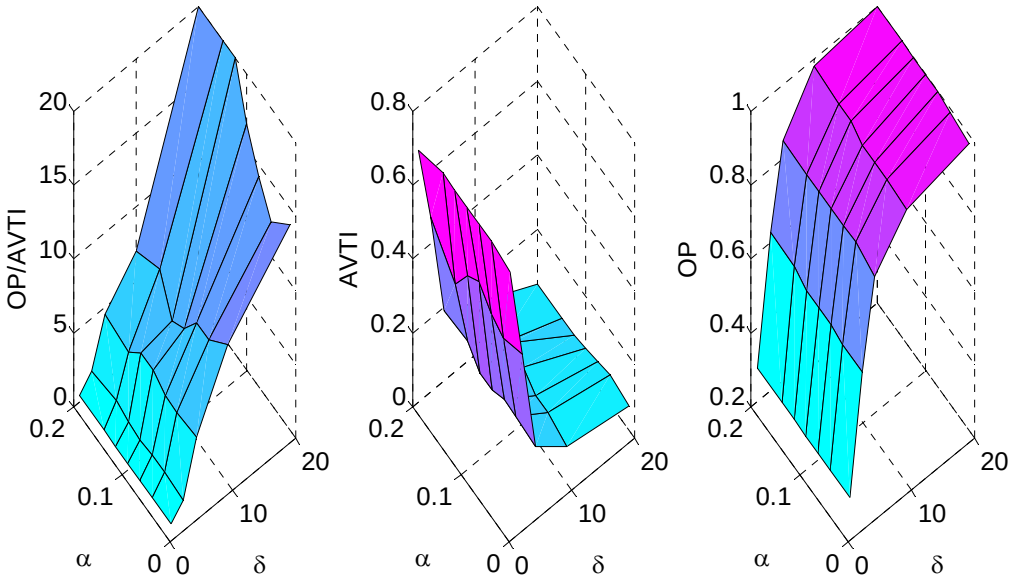

Figura 6.20: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para MTNT.

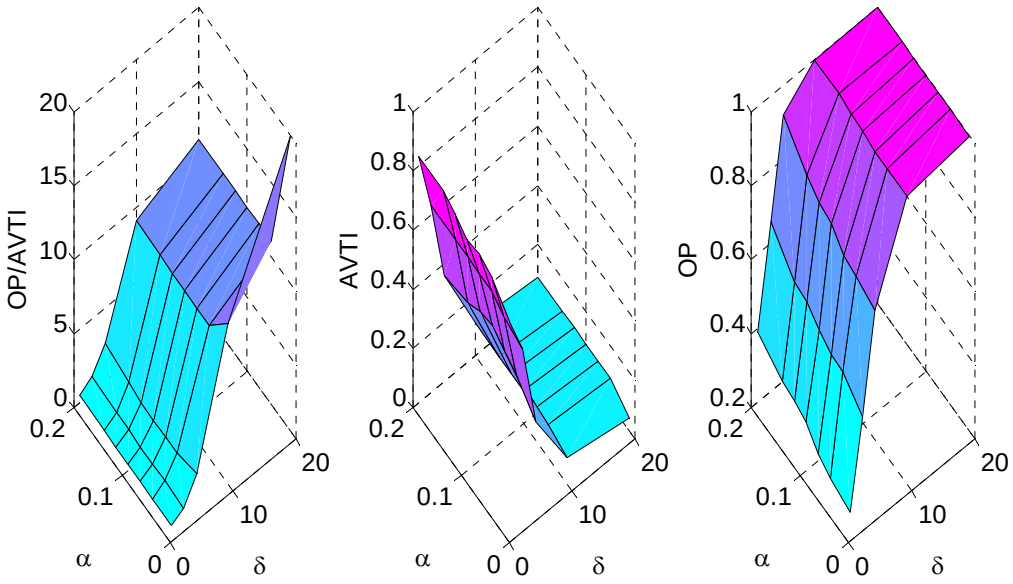

Figura 6.21: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para NTMT.

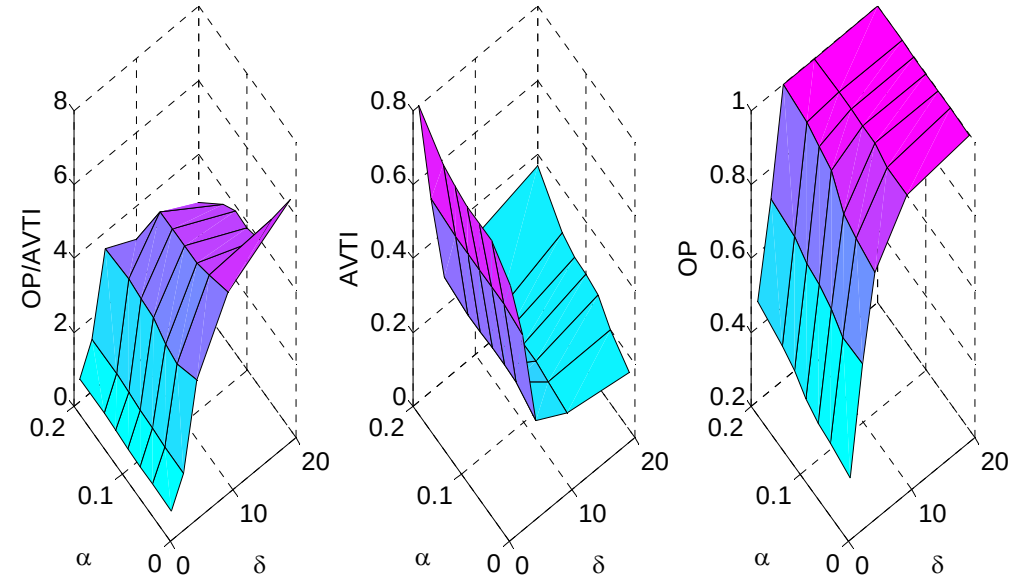

Figura 6.22: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para LCT.

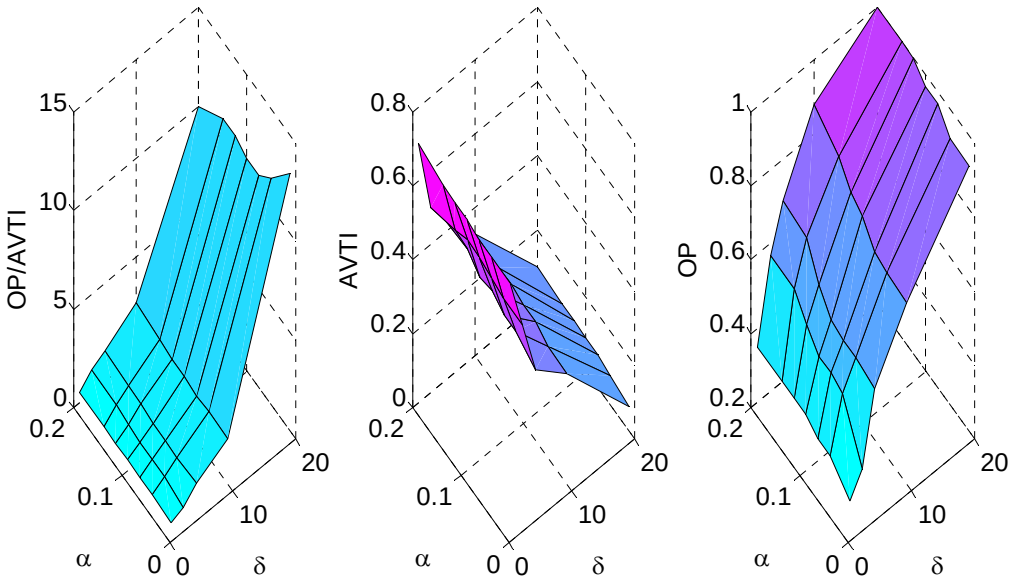

Figura 6.23: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para SEGE.

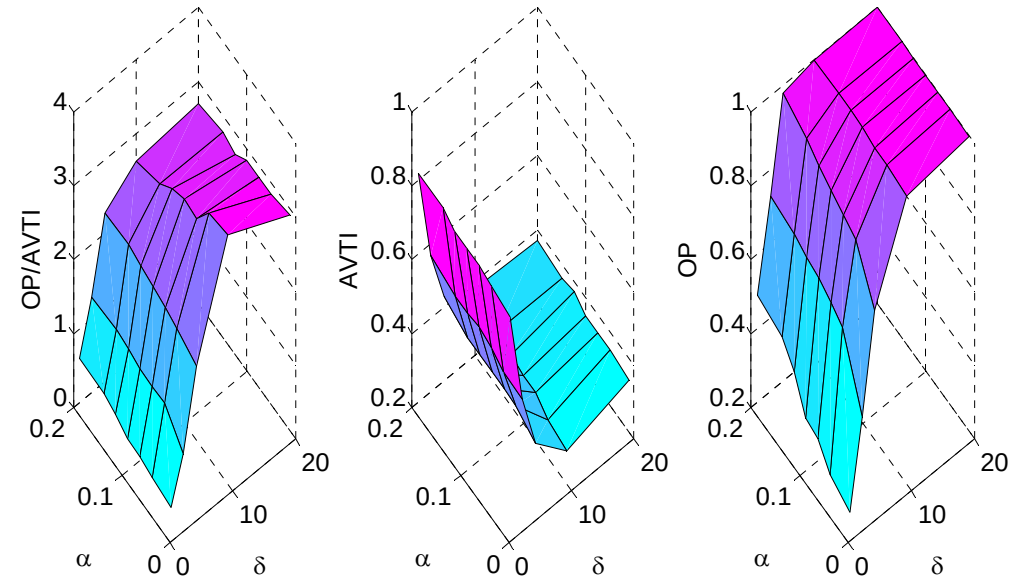

Figura 6.24: Curvas de OP/AVTI, OP e AVTI em função do tamanho do EG (δ) e nível de confiança (α) para SSCS.

Observando a seqüência de resultados vê-se que os algoritmos são pouco influenciados pelo nível de confiança, mas em alguns casos pode ser significativa. Por exemplo, no caso do SSCS, para um erro grosseiro de 10σ, a diferença entre utilizar um α de 0,01 ou de 0,2 geram uma diferença de OP de aproximadamente 10%.

A única estratégia que apresentam um comportamento mais diferenciado é o SEGE, mas isto pode ser relacionado à definição do intervalo de confiança que não está adequada

para que este possa ser comparado com os outros. Também se pode afirmar que o LCT se assemelha mais ao NTMT. Isto pode estar relacionado com a etapa de detecção que é realizada pelo NT nas duas estratégias. O mesmo aparece com o IMT e o MTNT e ambos têm em comum a detecção realizada pelo MT.

Estas curvas não foram encontradas na literatura para que se pudesse fazer qualquer comparação, apesar de trabalhos extensos realizados para o mesmo estudo de caso em Sanchez et al. (1999), Rollins e Davis (1992), Devanathanet al. (2004) e Romagnoli et al. (2000). Em Sanchez et al. (1999) existem alguns dados que poderiam ser comparados, mas os autores não deixam claro o tamanho do erro grosseiro, e geram uma confusão com os valores de α citando dois valores em momentos diferentes.

6.2.3 Sintonia das estratégias de detecção

Como já dito anteriormente, a comparação entre os algoritmos deve ser realizada na mesma base e esta não pode ser definida pela utilização de um mesmo intervalo de confiança. Esta foi realizada por tentativa e erro buscando valores de AVTI igual a 0,1 sob hipótese nula. Os valores para α são apresentados na Tabela 6.11.

Tabela 6.11: Graus de confiança obtidos por tentativa e erro

Algoritmo	1 - α	Algoritmo	1- α
SEGE	0,90	**MTNT**	0,87 e 0,77
LCT	0,84	**IMT**	0,87
SSCS	0,875	**NTMT**	0,87 e 0,77

Nem todos estes valores estão disponíveis na literatura. Para os disponíveis (SEGE, LCT, IMT, MIMT), os valores obtidos são semelhantes, comprovando que a metodologia para a sintonia está de acordo com o publicado.

6.2.4 Simulações Determinísticas

Como já explicitado no capítulo anterior, as simulações determinísticas são aquelas que não levam em consideração a presença de erro aleatório nos dados de processo e equivalem ao melhor desempenho possível dos algoritmos de detecção (Rollins et al., 1996; Sanchez et al., 1998). Para verificar qual é o melhor desempenho esperado para os algoritmos, foram realizadas simulações para três situações com diferentes quantidades de erros grosseiros simultâneos: 1, 2 e 3. Nestas simulações utilizou-se como tamanho de erro grosseiro, respectivamente: [10], [10, 20] e [10, 20, 30] desvios padrões (lembrando que para este estudo de caso 40 desvios padrões equivalem a 100% do valor da variável). As magnitudes dos erros grosseiros foram escolhidas para que se obtivesse uma boa idéia de qual é o valor para um alto poder de detecção.

No caso onde 3 erros grosseiros são introduzidos simultaneamente, é realizada a avaliação sem considerar a teoria dos erros equivalentes (chamada *3a*) e considerando-a (chamada *3b*). Os dados são apresentados em função dos indicadores de desempenho. Quando

considerada a *TEeq*, os indicadores de desempenho são corrigidos para contar como *acerto* caso o algoritmo tenha identificado um dos conjuntos equivalentes ao que está sendo simulado. Estes conjuntos já foram explicitados no Capítulo 5. Os resultados para Poder de Detecção Global (OP) estão relacionados na Tabela 6.12.

Tabela 6.12: OP em função do número de EGs para as simulações determinísticas

	1	**2**	**3a**	**3b**
LCT	1	0.805	0.80	0.91
IMT	1	0.950	0.90	0.934
MIMT	1	0.950	0.90	0.934
SSCS	1	0.881	0.68	0.78
SEGE	1	0.452	0.50	0.70
NTMT	1	0.738	0.31	0.47
MTNT	1	0.689	0.31	0.44

Para 3 erros grosseiros, os algoritmos IMT e MIMT apresentam um bom desempenho assim como o LCT e o SSCS. Utilizando a teoria dos erros equivalentes, os percentuais de acerto aumentam consideravelmente. Na Tabela 6.13 são apresentados os resultados para a Média de Erros Tipo I (AVTI). Dentre os algoritmos com alto poder de detecção, o IMT e o MIMT são os que apresentam os mais baixos AVTI. Para os algoritmos com poder de detecção mais baixo, o AVTI também o é, e este é o comportamento esperado, como já mencionado na introdução do Capítulo 4. Este é o caso para os algoritmos SEGE, NTMT e MTNT. Para avaliar a qualidade da estimação, são apresentados na Tabela 6.14 os resultados para Redução Total do Erro (TER).

Tabela 6.13: AVTI em função do número de EGs para as simulações determinísticas

	Número de EGs			
	1	**2**	**3a**	**3b**
LCT	0	0.952	1.4	1.4
IMT	0	0.54	0.58	0.47
MIMT	0	0.54	0.58	0.47
SSCS	0	2.095	2.3	1.2
SEGE	0	0.095	0.47	0.3
NTMT	0	0.476	0.78	0.6
MTNT	0	0.6190	0.75	0.5

Tabela 6.14: TER em função do número de EGs para as simulações determinísticas

	Número de EGs		
	1	**2**	**3**
LCT	1	0.67	0.57
IMT	1	0.75	0.57
MIMT	1	0.75	0.57
SSCS	1	0.74	0.31
SEGE	1	0.4	0.30
NTMT	1	0.67	0.43
MTNT	1	0.47	0.31

Como esperado, os algoritmos com maior poder de detecção apresentam melhor índice de TER. Isto se deve ao fato de que, se *TODOS* os erros grosseiros forem detectados corretamente, é garantido que os estimadores utilizados são *unbiased*. Como não foi adicionado erro aleatório, foi possível atingir um patamar de TER igual a 1 (valor desejado,

mas em geral não atingido), pois os erros aleatórios muitas vezes se cancelam e não podem ser estimados corretamente.Os dados relativos ao poder de perfeita identificação (OPF), para todos os algoritmos, são apresentados na Tabela 6.15.

Tabela 6.15: OPF em função do número de EGs para as simulações determinísticas

	Número de EGs		
	1	**2**	**3**
LCT	1.00	0.417	0.343
IMT	1.00	0.710	0.600
MIMT	1.00	0.710	0.600
SSCS	1.00	0	0.140
SEGE	1.00	0.435	0
NTMT	1.00	0.570	0.05
MTNT	1.00	0.403	0.029

Todos os algoritmos obtiveram desempenho ótimo na presença de um único erro grosseiro. Os algoritmos IMT e MIMT apresentam alto índice de perfeita identificação para todos os casos simulados (1, 2 e 3 erros grosseiros). Por ordem de desempenho, os métodos LCT, NTMT, MTNT apresentam um desempenho regular. O algoritmo SEGE apresentou pior desempenho quando na presença de 2 erros grosseiros devido ao efeito de cancelamento dos erros grosseiros já reportado em Rollins et al. (1996). Com o aumento do número de erros grosseiros, espera-se que os resultados para os algoritmos combinados sejam inferiores, pois estes utilizam o NT que apresenta a desvantagem do efeito de cancelamento dos erros grosseiros.

6.2.5 Simulações Monte Carlo

Para avaliar o comportamento dos algoritmos na presença do ruído de processo foram então realizadas simulações Monte Carlo. Os resultados para os indicadores *OPF*, *OP*, *AVTI* e *TER* são apresentados nas Tabelas 6.16, 6.17, 6.18, 6.19 e 6.20. Além disto, é apresentado o tempo computacional (t), dado em segundos. Nas primeiras três tabelas são mostrados os resultados em função do número de erros grosseiros. Já nas Tabelas 6.19 e 6.20 são apresentados os resultados totais médios em função da maneira com que é considerada a probabilidade de ocorrência: igual probabilidade de ocorrer 1, 2 ou 3 erros grosseiros ou igual probabilidade de ocorrer 1 dos 63 casos.

Ressalta-se que o procedimento para escolher a localização do erro grosseiro utilizado nesta dissertação é diferente do encontrado na literatura. Em geral, a localização é escolhida aleatoriamente, com igual probabilidade. Isto faz com que exista menor possibilidade de ocorrerem combinações com 2 erros grosseiros (e menor ainda com 3 erros grosseiros). Por este motivo, para comparação, são apresentados na Tabela 6.20 os valores totais equivalentes aos encontrados na literatura. Esta última é obtida a partir da média ponderada pelo número de combinações existentes em cada um dos 3 casos.

Tabela 6.16: Indicadores obtidos nas simulações Monte Carlo com um EG.

	OPF	OP	AVTI	TER	t
LCT	0.645	0.966	0.617	0.983	0.023
IMT	0.970	1.0	0.074	0.972	0.024
MIMT	0.970	1.0	0.074	0.972	0.024
SSCS	0.646	1.0	0.879	0.375	0.016
SEGE	0.999	0.999	0.001	0.401	0.031
NTMT	0.984	1.0	0.016	0.9555	0.042
MTNT	0.700	0.871	0.4084	0.8232	0.046

Na Tabela 6.16 nota-se que a estratégia com melhor desempenho na detecção é a SEGE. Já para a etapa de estimação o melhor desempenho é o *LCT*. Este comportamento é esperado e citado em Rollins et al. (1996) e Sanchez et al. (1999). O algoritmo SEGE utiliza uma estratégia combinatorial, avaliando a combinação que gera a menor função objetivo, reduzindo o cometimento de erro tipo I em relação às outras estratégias. Já a LCT é baseada no conceito da *Unbiased Estimation Technique*, que visa melhorar a etapa de estimação. Na Tabela 6.17 são apresentados os resultados para 2 erros grosseiros simultâneos no conjunto de dados.

Tabela 6.17: Indicadores obtidos nas simulações Monte Carlo com dois EGs.

	OPF	OP	AVTI	TER	t
LCT	0.19	0.795	1.187	0.540	0.010
IMT	0.71	0.757	0.064	0.719	0.005
MIMT	0.71	0.757	0.064	0.719	0.005
SSCS	0	0.915	2.079	0.464	0.020
SEGE	0.18	0.452	0.096	0.305	0.031
NTMT	0.29	0.768	0.469	0.698	0.012
MTNT	0.27	0.353	0.8116	0.475	0.011

Como esperado os algoritmos SEGE e LCT apresentam desempenhos bem inferiores aos possíveis (baseados na solução determinística). Este problema ocorre na presença de dois erros grosseiros e é reportado na literatura (Rollins et al., 1992 e Sanchez et al., 1996). Os autores relatam a dificuldade de detecção dos algoritmos SEGE e LCT para lidar com 11 dos 21 casos. Nestes casos não é obtida perfeita identificação. Por este motivo, em Sanchez et al. (1996), são realizadas uma série de modificações nos dois algoritmos recebendo novos nomes – MSEGE e MUBET. Estas modificações são baseadas na exclusão dos casos que falham e na modificação da topologia do estudo de caso para que seja atingido um patamar de OPF próximo a 1. O autor desta dissertação não concorda com a escolha de casos a serem simulados, pois os conjuntos que não apresentavam bons resultados não foram considerados após as modificações. Com isto obteve-se melhoria questionável dos resultados. Por este motivo estas modificações não foram implementadas.

Os algoritmos com melhor desempenho nas simulações Monte Carlo são o IMT e MIMT, levando em consideração a relação entre OP, AVTI e a etapa de estimação relacionada ao índice TER. O SSCS apresenta um poder de detecção maior, mas também apresentam um AVTI maior, e conseqüentemente apresenta TER menor. Este resultado confirma o apresentado por Rollins et al. (1998). Neste é realizado um estudo comparativo entre o SSCS e o LCT, apresentando resultados provenientes de simulações determinísticas e estocásticas e o efeito do pré-tratamento nos resultados. Com base nestes resultados é

demonstrado que, na presença de mais de um erro grosseiro, os algoritmos compensatórios do tipo do SSCS apresentam alto AVTI. Isto acontece especialmente quando a relação entre o tamanho do erro grosseiro e o ruído é grande. Além disto, os autores afirmam que embora o AVTI seja alto nesta condição, o algoritmo pode apresentar um desempenho muito superior a outras estratégias em outras condições. Estes resultados são condizentes com os resultados das simulações determinísticas.

Tabela 6.18: Indicadores obtidos nas simulações Monte Carlo com três EGs.

	OPF	OP	AVTI	TER	t
LCT	0.128	0.574	1.221	0.295	0.044
IMT	0.178	0.656	0.919	0.449	0.016
MIMT	0.178	0.656	0.919	0.449	0.016
SSCS	0.009	0.736	1.719	0.291	0.087
SEGE	0	0.184	1.221	0.389	0.033
NTMT	0	0.571	1.114	0.443	0.053
MTNT	0	0.567	1.0951	0.444	0.051

Para a presença de 3 erros grosseiros o desempenho dos algoritmos é bem inferior aos obtidos na solução determinística. Provavelmente o poder de detecção é dependente da localização e, além disto, já foi demonstrado pela análise de observabilidade que existe uma redução de 60% na ajustabilidade e de 50% na observabilidade, em relação ao caso sem erros grosseiros. Isto está de acordo com o apresentado na Tabela 6.18. Nas Tabelas 6.19 e 6.20 são apresentados os resultados médios, ponderados de maneira diferente.

Tabela 6.19: Resultados com igual probabilidade dos 63 casos ocorrerem.

	OPF	OP	AVTI	TER	t
LCT	0.136	0.728	1.018	0.591	0.013
IMT	0.199	0.691	1.143	0.453	0.030
MIMT	0.136	0.728	1.018	0.591	0.013
SSCS	0.077	0.825	1.746	0.358	0.057
SEGE	0.111	0.364	0.710	0.362	0.032
NTMT	0.208	0.731	0.669	0.588	0.018
MTNT	0.078	0.529	0.924	0.497	0.037

Tabela 6.20: Resultados com igual probabilidade de ocorrer 1, 2 ou 3 EGs.

	OPF	OP	AVTI	TER	t
LCT	0.096	0.804	1.119	0.697	0.015
IMT	0.314	0.778	1.008	0.606	0.026
MIMT	0.096	0.804	1.119	0.697	0.015
SSCS	0.218	0.884	1.559	0.377	0.041
SEGE	0.333	0.545	0.439	0.365	0.032
NTMT	0.387	0.808	0.468	0.701	0.023
MTNT	0.233	0.597	0.772	0.581	0.036

Para comparação são mostrados todos os resultados para TER e OP nas Figuras 6.25 e 6.26. Sob ponto de vista de detecção, o algoritmo SSCS apresentou melhor desempenho nas 3 situações apresentadas. Já com base no conjunto reconciliação + estratégia de detecção, os algoritmos IMT e MIMT apresentaram os melhores resultados globais. Por este motivo o IMT foi escolhido para ser modificado gerando o algoritmo IMT robusto.

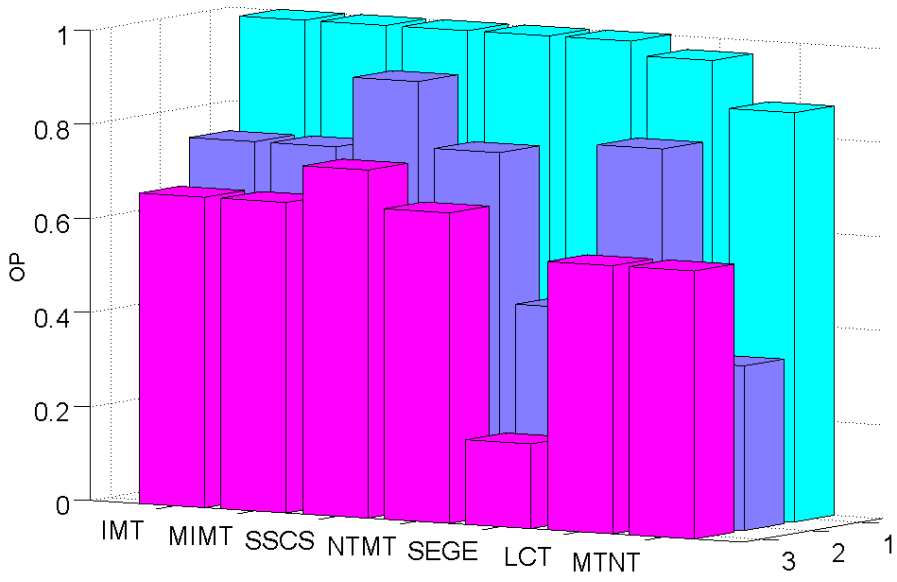

Figura 6.25: Poder de detecção global nas simulações Monte Carlo.

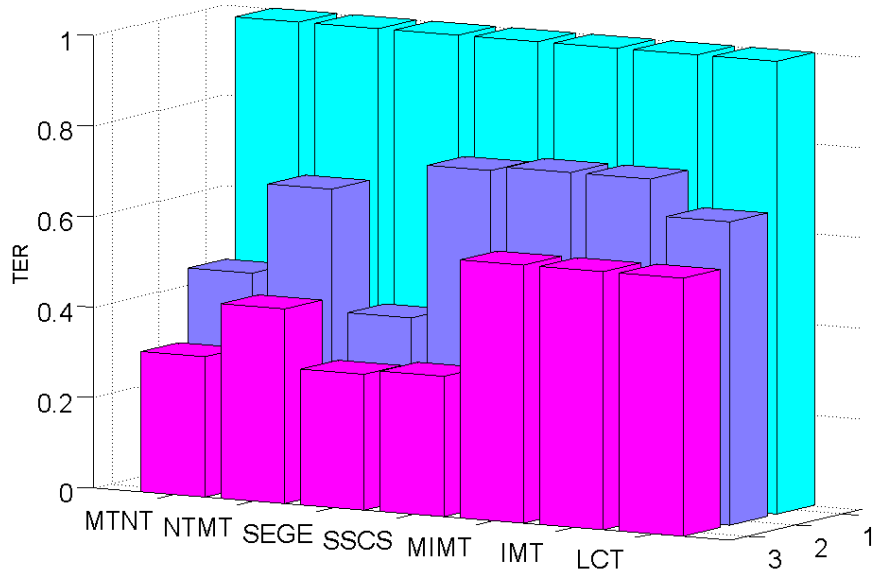

Figura 6.26: Redução Total do Erro nas simulações Monte Carlo.

6.2.6 Influência do pré-tratamento dos dados

Para verificar a influência do pré-tratamento, foi utilizado o mesmo procedimento aplicado no item 6.1.2 . Todos os algoritmos foram aplicados nos dados filtrados, da mesma forma aplicada nas simulações Monte Carlo, para que pudesse ser realizada comparação. Na Tabela 6.21 são apresentados os resultados para um erro grosseiro para aplicação do filtro FIR (os filtros apresentaram comportamento similar, por isto só serão apresentados estes resultados). Nas Figuras 6.27 e 6.28 são comparados os resultados das simulações Monte Carlo para *OPF* e *TER,* respectivamente, com os resultados obtidos pelos dados pré-filtrados.

Tabela 6.21: Resultados obtidos em simulações Monte Carlo com dados filtrados

	OPF	OP	AVTI	TER	t
LCT	1.000	1.000	0.000	0.993	0.004
IMT	0.99	1.000	0.002	0.988	0.005
MIMT	0. 99	1.000	0.002	0.988	0.005
SSCS	1.000	1.000	0.000	0.970	0.010
SEGE	1.000	1.000	0.000	0.978	0.030
NTMT	1.000	1.000	0.000	0.993	0.007
MTNT	0.857	0.857	0.286	0.956	0.005

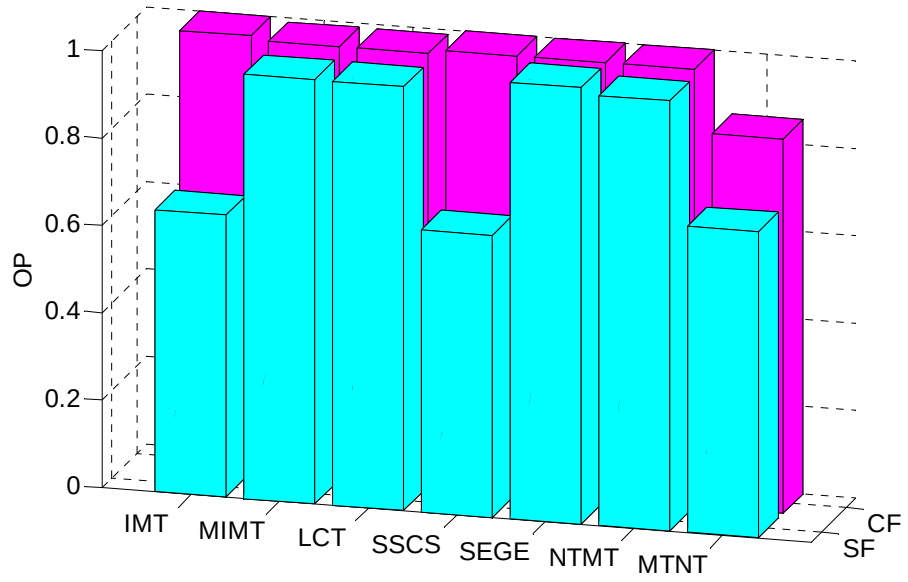

Figura 6.27: OPF para dados filtrados (CF) e brutos (SF).

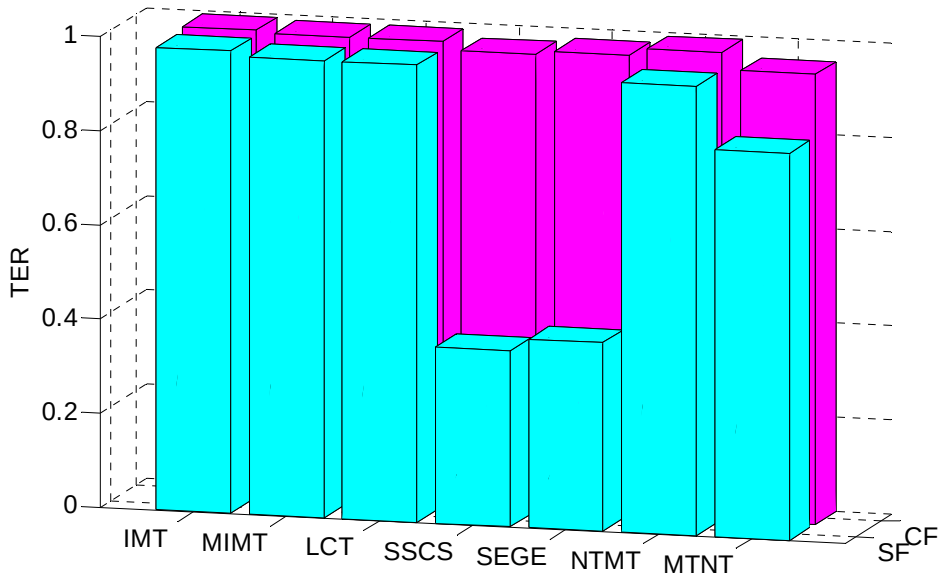

Figura 6.28: TER para dados filtrados (CF) e brutos (SF).

Observa-se que todos os algoritmos apresentaram desempenho muito próximo do obtido nas simulações determinísticas. Para os algoritmos que não obtinham uma alta performance os resultados demonstram que o pré-tratamento dos dados é realmente eficiente para ser utilizado junto com as estratégias de detecção de erros grosseiros. Não foram realizados os testes completos para comparar todos os resultados obtidos visto que estudos Monte Carlo são longos. Mas pretende-se continuar nesta linha de pesquisa em trabalhos futuros.

6.2.7 Comparação com a reconciliação robusta

Para iniciar a comparação com a reconciliação robusta, as diferentes funções objetivo foram submetidas a dois conjuntos de simulações determinísticas. O primeiro visando à comparação entre os métodos com a solução tradicional e o segundo conjunto de simulações para a comparação com os resultados já apresentados no item 6.2.5, para as estratégias de detecção de erros grosseiros. O objetivo do primeiro conjunto de rodadas é comparar a solução com as funções robustas frente à solução com a função tradicional de mínimos quadrados ponderados. Para isto simulou-se o conjunto de dados na presença de 1, 2 ou 3 erros grosseiros e variou-se o tamanho do erro grosseiro entre 1 - 40 desvios padrões. Foram simulados todos os 63 casos já apresentados anteriormente, considerando todas as combinações possíveis de variáveis com erros grosseiros. Os resultados obtidos na presença de 1 erro grosseiro são apresentados na Tabela 6.22.

Tabela 6.22: Resultados para TER obtidos nas simulações determinísticas com um EG

	Tamanho do EG simulado					
Função Objetivo	**1**	**5**	**10**	**20**	**30**	**40**
Normal (SQP)	0.38	0.38	0.38	0.38	0.38	0.38
Normal Contaminada(CN)	0.38	0.72	0.98	0.99	0.97	0.95
Cauchy (CA)	0.33	0.67	0.61	0.79	0.69	0.71
Fair (FA)	0.44	0.57	0.64	0.69	0.73	0.74

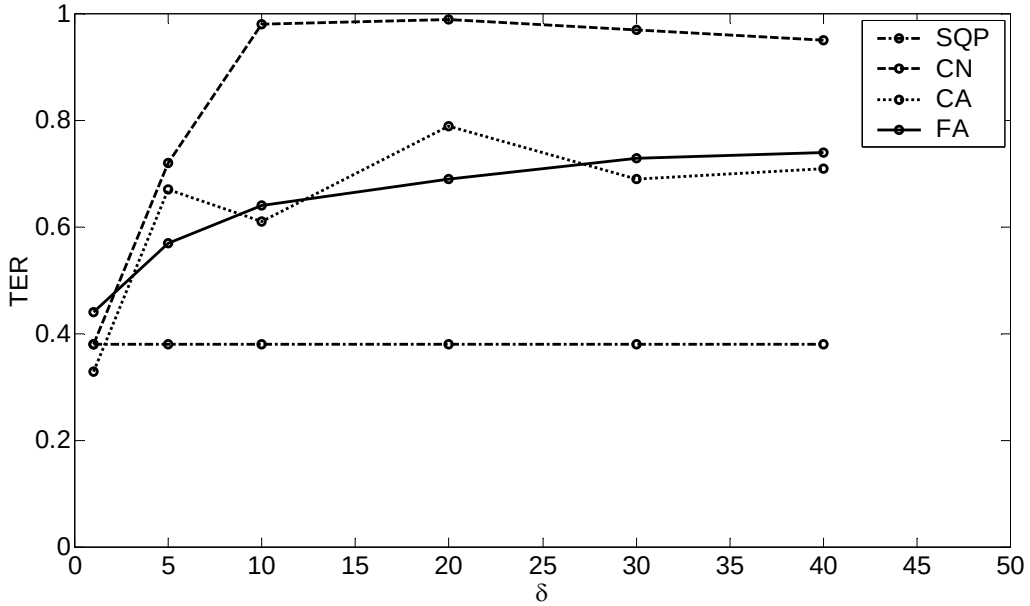

Figura 6.29: Resultados apresentados na Tabela 6.22 para *TER*.

Na Figura 6.29 observa-se que para um erro grosseiro de um desvio padrão, os estimadores apresentam praticamente o mesmo desempenho da função normal tradicional. Isto é esperado visto que as funções objetivo aproximam-se da função normal neste intervalo. À medida que o tamanho do erro grosseiro aumenta, os estimadores passam a ter um comportamento diferenciado e os dois estimadores com melhor resultado são os baseados na função objetivo normal contaminada e na função objetivo de Cauchy.Os resultados obtidos para 2 erros grosseiros são apresentados na Tabela 6.23 e demonstrados na Figura **6.30**.

Tabela 6.23: Resultados para TER obtidos nas simulações determinísticas com dois EGs

	Tamanho do EG simulado					
Função Objetivo	**1**	**5**	**10**	**20**	**30**	**40**
SQP	0.28	0.28	0.28	0.28	0.28	0.29
CN	0.29	0.72	0.76	0.87	0.72	0.89
CA	0.33	0.53	0.72	0.79	0.65	0.79
FA	0.31	0.37	0.41	0.48	0.47	0.54

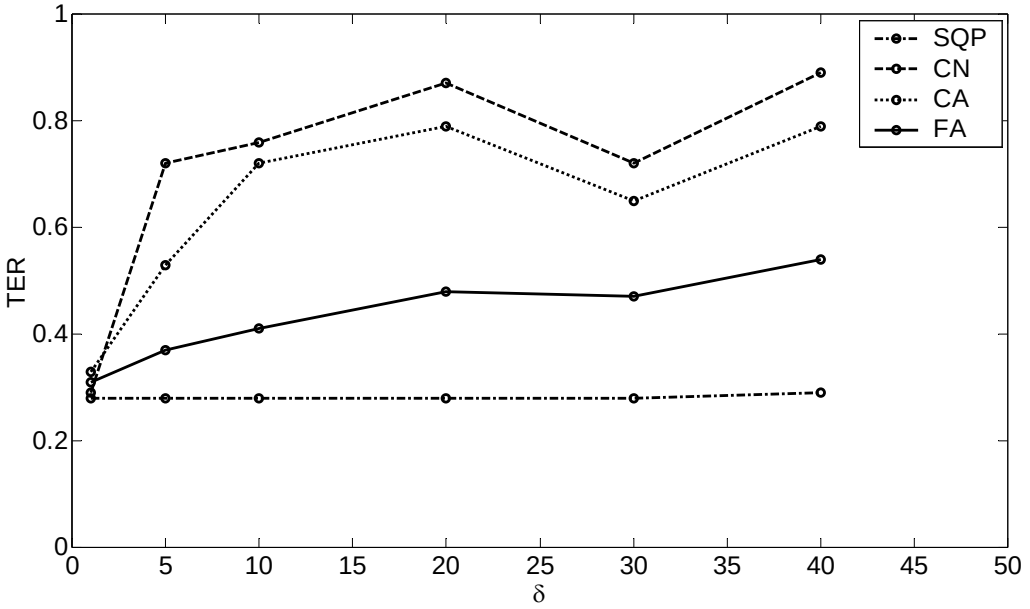

Figura 6.30: Resultados apresentados na Tabela 6.23 para *TER.*

Com base na Figura 6.30 fica evidente que a função objetivo que apresenta a melhor estimação é a Normal Contaminada. Como esperado, quanto maior o erro grosseiro, melhor é o desempenho das funções robustas e o aumento no número de erros grosseiros diminui um pouco o indicador TER. Isto se deve à redundância de menor qualidade quando comparado com o sistema na presença de um único erro grosseiro. Na Tabela 6.24 e na Figura 6.31 são apresentados os resultados para as simulações na presença de três erros grosseiros. Este é o máximo de erros que podem ser estimados para garantir que haja alguma reconciliação nos dados. Da mesma forma que acontece na presença de um e dois erros grosseiros, a função normal contaminada apresenta o melhor desempenho para três erros grosseiros nas simulações determinísticas. A forma das curvas está de acordo com a esperada, baseada nos conceitos apresentados no Capítulo 3, mostrando que as funções estão sintonizadas de maneira que possam ser comparadas de maneira justa.

Tabela 6.24: Resultados para TER obtidos nas simulações determinísticas com três EGs

	Tamanho do EG simulado					
Função Objetivo	**1**	**5**	**10**	**20**	**30**	**40**
SQP	0.22	0.22	0.22	0.22	0.22	0.22
CN	0.22	0.23	0.30	0.34	0.35	0.44
CA	0.23	0.24	0.28	0.32	0.32	0.34
FA	0.22	0.22	0.22	0.22	0.23	0.23

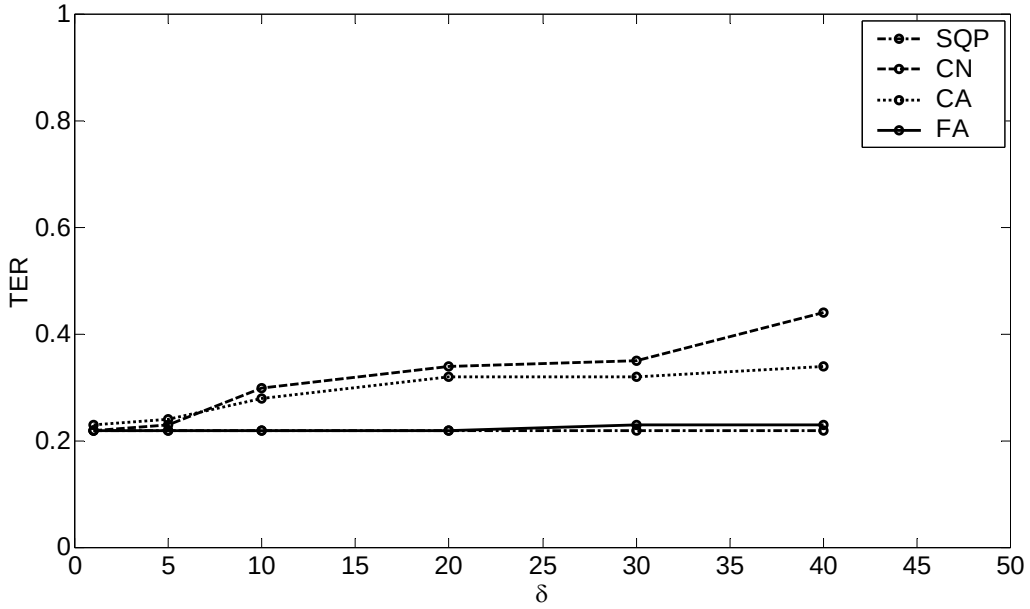

Figura 6.31: Resultados apresentados na Tabela 6.24 para *TER*.

Na Tabela 6.25 e na Figura 6.32 é exibido um resumo com os resultados obtidos para as estratégias convencionais e para a reconciliação robusta. Esta é dada em função da quantidade de erros grosseiros, para condições equivalentes às simulações determinísticas realizadas.

Tabela 6.25: Resumo dos resultados obtidos para TER nas simulações determinísticas

	1	**2**	**3**
LCT	1	0.67	0.57
IMT	1	0.75	0.41
MIMT	1	0.75	0.41
SSCS	1	0.74	0.31
SEGE	1	0.40	0.30
NTMT	1	0.67	0.43
MTNT	1	0.47	0.31
Gaussiana (Norm.)	0.38	0.28	0.22
Normal Contaminada (CN)	0.98	0.50	0.33
Cauchy (CA)	0.61	0.48	0.32
Fair (FA)	0.64	0.52	0.31

Apesar de a reconciliação robusta ser, de certa maneira, eficiente para tratar dados com erros grosseiros sem a necessidade de se utilizar nenhuma outra rotina, esta apresenta a desvantagem de não localizar o erro grosseiro. Quando comparada com os resultados obtidos pelas estratégias de detecção, estas últimas ainda apresentam características mais interessantes. O fato de estas estratégias conseguirem eliminar o erro quase que completamente e possuírem a etapa de detecção/localização dos erros grosseiros são duas vantagens atrativas. A desvantagem das estratégias é o cometimento de erros tipo I devido ao efeito "smearing".

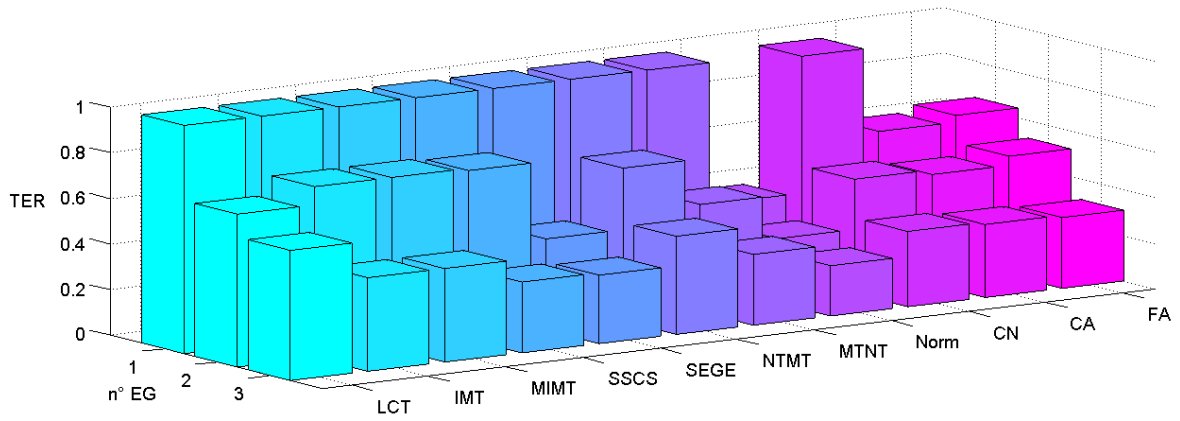

Figura 6.32: Comparação entre as estratégias de detecção e a reconciliação robusta - TER.

Na comparação pode-se ver que os reconciliadores robustos obtiveram um bom resultado frente aos algoritmos de detecção de erros grosseiros obtendo resultado até superior que algumas das estratégias de detecção quando na presença de 2 erros grosseiros. É interessante mostrar que, mesmo na presença de um erro grosseiro - situação em que os métodos de detecção apresentam um resultado ideal, as estimativas obtidas pelas funções robustas obtiveram um alto índice de acerto. Entre as funções objetivo a que apresentou melhor resultado foi a Normal Contaminada, seguida da função de Cauchy.

E neste momento é apresentada a justificativa para a escolha da função objetivo normal contaminada para a implementação do novo método de detecção de erros grosseiros IMT-robusto. A idéia é utilizar um dos melhores métodos de detecção de erros grosseiros (e com certeza o mais simples) aliado a função objetivo que apresentou os melhores resultados. Como o método IMT é bem simples, é possível substituir a função objetivo por qualquer outra que se queira, sem grandes problemas de adaptação.

6.3 Desenvolvimento e Validação do IMT robusto

Esta seção será dividida em duas partes principais. A primeira tem por objetivo definir a filosofia de detecção mais adequada, visto que as duas opções parecem interessantes à primeira vista. Para isto será utilizado o estudo de caso 1, já exaustivamente explorado neste trabalho. Foram realizadas simulações determinísticas e simulações Monte Carlo, nos moldes dos resultados apresentados nos itens anteriores, e os resultados são comparados. Estes

primeiros resultados já validariam o método, mas também são apresentados resultados para o Estudo de Caso 2. Estes foram gerados com base em uma publicação encontrada na literatura e servem para comparar resultados, validar o método e identificar possíveis melhorias.

6.3.1 Estudo de Caso 1: Simulações Determinísticas

Este estudo foi realizado com as mesmas condições já apresentadas no item 6.2.2. Os dados são apresentados em função dos indicadores de desempenho apresentados no Capítulo 4. Salienta-se que o algoritmo IMTr1 é a implementação do IMT somente mudando a função objetivo da reconciliação. Já o IMTr2 refere-se à identificação simultânea de todos os erros grosseiros. Para o IMTr2 o número máximo de erros grosseiros foi estabelecido como sendo 3, visto que seria injusto fazer com que o algoritmo acertasse a quantidade de erros sem nenhuma avaliação prévia. Os resultados relativos ao poder de perfeita identificação (OPF) são apresentados na Tabela 6.26. Na Tabela 6.27 são apresentados os resultados para Poder de Detecção Global (OP).

Tabela 6.26: OPF obtidos nas simulações determinísticas.

	N° de Erros Grosseiros		
	1	**2**	**3**
IMT	1	0.71	0.60
IMTr	1	0.85	0.71
IMTr2	1	0.65	0.60

Tabela 6.27: OP obtidos nas simulações determinísticas.

	N° de Erros Grosseiros		
	1	**2**	**3**
IMT	1	0.95	0.90
IMTr1	1	0.97	0.95
IMTr2	1	0.76	0.69

Pode-se notar que o algoritmo IMTr1 apresenta resultado superior para perfeita identificação em todos os casos e resultados inclusive superiores à versão original. O IMTr2 não apresenta resultados tão bons quando aumenta o número de erros grosseiros no sistema. Provavelmente isto se deve ao que foi publicado por Bagajewicz (2007) sobre a falta de consistência estatística do teste *MT* na presença de mais de um erro grosseiro. Neste caso o autor afirma que o teste continua sendo um bom indicativo para ser utilizado em estratégias iterativas, mas somente a medição que mais excede o valor crítico do teste estatístico tem sentido. Na Tabela 6.28 são apresentados os resultados para Média de Erros Tipo I (AVTI) e na Tabela 6.29 estão os resultados obtidos para TER.

Tabela 6.28: AVTI obtido nas simulações determinísticas.

	N° de Erros Grosseiros		
	1	**2**	**3**
IMT	0	0.54	0.47
IMTr1	0	0.10	0.50
IMTr2	0	0.87	0.95

Tabela 6.29: TER obtido nas simulações determinísticas.

	N° de Erros Grosseiros		
	1	**2**	**3**
IMT	1	0.75	0.57
IMTr1	1	0.98	0.70
IMTr2	0.76	0.70	0.53

O que mais chama atenção nestes resultados é o valor pequeno de AVTI para o algoritmo IMTr1 e o alto índice de redução do erro, mesmo para um poder de detecção menor que 1 (caso com 2 erros grosseiros) inclusive quando comparado pelos valores obtidos pelos outros métodos de detecção (apresentados na seção 6.2.5). Este bom desempenho será confirmado na próxima seção, onde são apresentadas as simulações Monte Carlo.

6.3.2 Estudo de Caso 1: simulações Monte Carlo

Foram realizadas simulações Monte Carlo de maneira semelhante ao já apresentado no item 6.2.5 . Os resultados serão comparados com a utilização do algoritmo IMT, visando uma comparação justa. Na Tabela 6.30 são apresentados os valores para detecção de um EG. Já na Tabela 6.31 são apresentados os valores para a detecção de um EG e a utilização do pré-tratamento dos dados, apresentado na seção 6.1.2.

Tabela 6.30: Resultados obtidos nas simulações Monte Carlo com um EG

	OPF	**OP**	**AVTI**	**TER**	**t**
IMT	1	1	0	1	0.024
IMTr1	0.909	0.917	0	1	0.210
IMTr2	0.710	0.730	0.48	0.94	0.305

Observa-se na Tabela 6.30 que para um erro grosseiro tanto os valores de OP e AVTI são inferiores à versão original, mas mesmo assim o IMTr1 consegue reduzir o erro nos dados a ponto de atingir TER igual a 1. A maior desvantagem em relação à versão original é o tempo computacional. A versão IMTr2 comete muitos erros do tipo I, visto que ela não escolhe um candidato, mas sim todos os candidatos que não passam no teste MT versão robusta.

Tabela 6.31: Resultados obtidos nas simulações Monte Carlo com um EG e filtragem

	OPF	**OP**	**AVTI**	**TER**	**t**
IMT	1	1	0	1	0.005
IMTr1	1	1	0.05	1	0.255
IMTr2	0.742	0.919	0.939	0.98	0.236

Com base nos resultados apresentados na Tabela 6.31 conclui-se que o pré-tratamento de dados também melhora o desempenho dos algoritmos em todos os indicadores. Na Tabela 6.32 estão os resultados para 2 EGs.

Tabela 6.32: Resultados obtidos nas simulações Monte Carlo com dois EGs

	OPF	**OP**	**AVTI**	**TER**	**t**
IMT	0.71	0.757	0.064	0.72	0.005
IMTr1	0.845	0.900	0.051	0.96	0.258
IMTr2	0.571	0.718	1.4	0.69	0.269

Na presença de dois erros grosseiros é que aparece a grande diferença entre as implementações. Para o IMTr1 os indicadores OP e OPF são bem superiores aos valores obtidos pelo IMTr2 devido ao alto AVTI obtido por este último. Além disto, o índice TER é muito superior, inclusive que a versão original, quase atingindo 1. Na Tabela 6.33 estão os resultados para três erros grosseiros.

Tabela 6.33: Resultados obtidos nas simulações Monte Carlo com três EGs

	OPF	OP	AVTI	TER	t
IMT	0.178	0.656	0.919	0.449	0.016
IMTr1	0.800	0.8763	0.129	0.688	0.189
IMTr2	0.685	0.718	0.573	0.502	0.347

Na presença de três erros grosseiros o desempenho do IMTr1 se distancia mais ainda dos outros algoritmos mostrando que definitivamente é superior. A habilidade de detecção é melhorada em relação ao algoritmo original, assim como um desempenho superior é apresentado para a redução do erro. O algoritmo IMTr1 ainda poderia ser melhorado com uma sintonia diferente em função da razão entre o tamanho do erro grosseiro e o ruído. Na Figura 6.33 são resumidos os desempenhos obtidos pelos algoritmos IMTr e IMT. Para isto foi realizados a média considerando igual probabilidade de ocorrer um, dois ou três erros grosseiros.

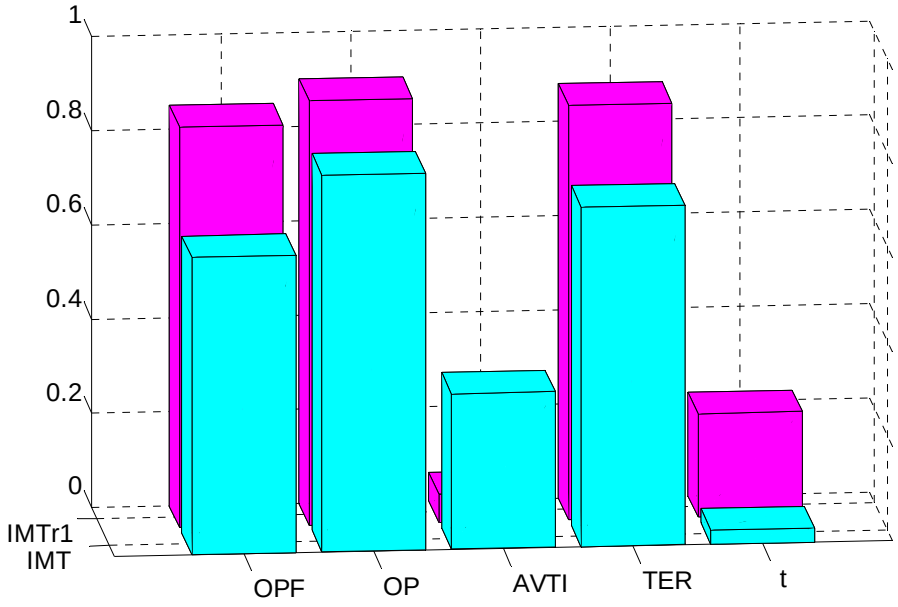

Figura 6.33: Resumo dos resultados obtidos nas simulações Monte Carlo para IMT e IMTr1.

6.3.3 Estudo de Caso 2

Para este estudo de caso foram reproduzidos os testes utilizados por Rollins et al. (1996) e Ozyurt e Pike (20004). Para comparar os resultados com os obtidos pelos dois artigos, será utilizado o caso onde existem 7 erros grosseiros. Os erros grosseiros foram gerados de maneira que pudesse ser qualquer valor entre 12,5% – 62,5% dos valores

verdadeiros já apresentados no capítulo anterior. A localização e a magnitude foram escolhidas aleatoriamente com igual probabilidade de ocorrerem.

Para avaliar a presença de 7 erros grosseiros foram testados os algoritmos IMT e IMTr1. Os resultados reportados por Ozyurt e Pike (2004) referentes ao MIMT e à solução do problema de otimização utilizando a função normal contaminada foram reproduzidos na Tabela 6.34 junto com os resultados obtidos pelas simulações Monte Carlo. O mesmo acontece com os resultados publicados por Rollins et al. (1996) para os algoritmos LCT e SSCS. Nesta são apresentados também o número de simulações rodadas para obtenção dos resultados visto que o algoritmo SQP apresentou problemas para resolver o problema de otimização.

Tabela 6.34: Resultados obtidos para o Estudo de caso 2.

	MIMT	**CN**	**IMT**	**IMTr**	**LCT**	**SSCS**
n	1000	1052	1000	2189	1000	1000
OP	0.684	0.724	0.517	0.787	0.257	0.36
AVTI	1.364	3.371	2.02	1.20	*	1.23

* não publicado

Observa-se que o IMTr apresenta desempenho superior, apesar dos problemas encontrados pela rotina SQP. Os dados provenientes de rodadas em que não houve convergência foram descartados. Para tentar superar este problema tentou-se escalonar o problema visto que este processo apresenta vazões com diferentes ordens de grandeza. Não foi obtido um bom resultado. O outro artifício tentado foi o proposto Rollins et al. (1996) de se retirar da análise as 9 menores vazões. Infelizmente este também falhou. Talvez a solução seja a utilização de outras técnicas para a solução do problema. Estas não foram exploradas até o momento, mas estão nos planos de trabalhos futuros.

Capítulo 7

Conclusões e Sugestões

7.1 Conclusões finais

O sistema de reconciliação de dados trata de um problema advindo da evolução das técnicas de medição, aquisição e armazenamento de dados. Esta evolução permitiu que dados de processo fossem armazenados em grande quantidade e o estado real de um processo fosse mais bem caracterizado. O sistema de reconciliação tem o papel de garantir a consistência destes dados, utilizando a redundância das variáveis medidas em conjunto com um modelo estatístico da medição para aumentar a precisão dos dados. O procedimento completo tem por objetivo que as equações de conservação sejam satisfeitas, tratando dos erros aleatórios inerentes ao processo de medição e que eventuais erros grosseiros sejam detectados e corrigidos. Estes duas últimas atribuições referem-se aos dois problemas tratados neste trabalho: avaliação de técnicas para solução do problema de reconciliação e para detecção de erros grosseiros.

Existem diferentes técnicas para se obter a solução do problema de reconciliação de dados. Tradicionalmente é resolvido um problema de otimização com uma função objetivo de mínimos quadrados ponderados. Nesta dissertação, esta técnica foi aplicada e avaliou-se a influência da presença de variáveis não-medidas e a utilização de pré-tratamento de dados. Os resultados mostraram que, dependendo da localização e do número de variáveis não medidas, algumas variáveis medidas passam a se comportar como não redundantes na prática. Nestas variáveis o ajuste realizado na reconciliação é ínfimo e isto compromete muito a qualidade dos valores estimados. Em contrapartida avaliou-se o pré-tratamento de dados, comumente utilizado na indústria, mas ainda não explorado na literatura de reconciliação de dados. Verificou-se que sua utilização pode gerar ganhos significativos para reconciliação de dados estacionária linear.

A solução tradicional do problema de reconciliação necessita que duas premissas sejam satisfeitas para manter a sua validade do ponto de vista estatístico: que o erro aleatório siga uma distribuição gaussiana e que não existam erros grosseiros nos dados. Caso existam erros grosseiros, estes serão espalhados por todo o conjunto de dados por causa de um efeito

conhecido como *smearing*, prejudicando inclusive as medidas livres de erros. Isto é uma desvantagem, já que nem sempre se podem garantir tais premissas. Por este motivo existe um número considerável de estratégias que visam detectar os erros grosseiros. Na detecção de erros grosseiros clássica são utilizados testes estatísticos, aproveitando as propriedades conhecidas das variáveis aleatórias gaussianas para cumprir três etapas: detecção, localização e estimação do erro grosseiro.

Foram avaliadas 6 estratégias populares na literatura, diferentes em sua concepção para uma ou mais etapas. As estratégias são: IMT (e MIMT), LCT, SSCS, SEGE, NTMT e MTNT. Nesta etapa de comparação entre as diferentes estratégias foi utilizado um estudo de caso clássico da literatura, escolhido devido ao seu tamanho reduzido e possibilidade de realização de um estudo combinatorial completo. O estudo de caso foi avaliado frente a influência da topologia na detecção de erros grosseiros, de maneira similar à realizada para a reconciliação de dados. Para avaliação e comparação das estratégias, foram levantadas as curvas de poder de detecção em função do tamanho do erro grosseiro e do nível de confiança dos testes estatísticos. Com base nos resultados destas curvas pode-se ver que os algoritmos necessitam que o nível de confiança seja ajustado, de modo que a comparação seja justa. Portanto, os algoritmos foram sintonizados de maneira que apresentassem AVTI igual a 0,1.

A comparação das estratégias foi realizada com base em um estudo combinatorial completo. Neste foram testadas todas as combinações possíveis de erros grosseiros e realizadas simulações determinísticas e simulações Monte Carlo. Como resultado obteve-se que, para este estudo, o algoritmo IMT é o mais eficiente quando consideradas as 3 etapas que devem ser cumpridas por uma estratégia de detecção. Já as estratégias NTMT e MTNT, que são pouco exploradas na literatura, apresentaram resultados surpreendentemente satisfatórios. O LCT, o SEGE e o SSCS tiveram bons resultados, mas algumas deficiências já reportadas na literatura puderam ser observadas. O pré-tratamento de dados também foi avaliado no contexto da detecção de erros grosseiros e apresentou resultados promissores, melhorando bastante a estimação dos estados.

Além das estratégias de detecção clássicas, existe outra possibilidade de tratamento de dados na presença de erros grosseiros. Podem-se utilizar, ao invés da função objetivo tradicional, outras famílias de funções objetivo baseadas em estatística robusta. Estas têm a habilidade de rejeitar a presença de grandes erros, diminuindo o efeito *smearing*. Neste contexto foram avaliadas 3 funções objetivo diferentes: a função objetivo normal contaminada, a função de Cauchy e a função Fair. Para comparação com as estratégias de detecção, as funções objetivo foram sujeitas ao mesmo conjunto de simulações determinísticas e os resultados comparados. Como resultado da comparação conclui-se que o desempenho na habilidade de rejeitar erros grosseiros é bom, entretanto as estratégias de detecção clássicas apresentam a vantagem das etapas de detecção e localização.

Com isto em mente, foi então proposto um novo algoritmo de detecção de erros grosseiros chamado de IMT robusto. Esta é contribuição principal deste trabalho. Neste propõe-se utilizar, ao invés da função objetivo normal, uma função objetivo robusta para a etapa de detecção e estimação do erro. São apresentadas as etapas de desenvolvimento, onde

foi escolhida a filosofia de detecção e a função objetivo robusta. A filosofia de detecção escolhida foi a mesma que a do IMT original, iterativa e eliminatória. A função objetivo com melhores resultados foi a normal contaminada. Para a validação do algoritmo foram realizados os mesmo conjuntos de simulações determinísticas e estocásticas apresentados para a avaliação das estratégias clássicas. O IMTr apresentou um bom desempenho geral e resultados superiores na etapa de estimação. Por fim, o algoritmo foi aplicado a um segundo estudo de caso e os resultados foram comparados aos reportados na literatura. O IMTr apresentou resultados superiores, mas como o estudo de caso utilizado era mais complexo, apareceram problemas para o solução do problema de otimização que não foram resolvidos.

7.2 Sugestões para trabalhos futuros

Como sugestões para trabalhos futuros pode-se citar:

1. Avaliar a utilização de outras funções objetivo robustas como, por exemplo, *Redescending Estimators*, Logística e Lorentzian.

2. Desenvolver uma metodologia de pré-tratamento de dados que possa ser incluída no problema de reconciliação, assim como para tratamento de *outliers* e detecção de estados estacionários.

3. Adaptar as outras estratégias de detecção à versão robusta.

4. Estender este estudo para problemas bilineares e não-lineares.

5. Realizar estudo semelhante em dados reais de processo.

6. Desenvolver uma metodologia para projetar redes de medição vislumbrando as necessidades de observabilidade, precisão mínima e custo de medição.

Referências Bibliográficas

Abu-el-zeet, Z. H., Becerra, V. M., Roberts, P. D., (2002), Combined bias and outlier identification in dynamic data reconciliation. *Computers and Chem. Engng.*, 26 (2), 921–935.

Albuquerque, J. S., Biegler, L. T., (1996), Data Reconciliation and Gross-Error Detection for Dynamic Systems, *AIChE Journal*, v. 42, pp. 2841-2856.

Alhaj–Dibo, M., Maquin, D., Ragot, J., (2008), Data Reconciliation: A Robust Approach Using a Contaminated Distribution, *Control Engineering Practice*, v. 16, pp. 159-170.

Ali, Y., Narashiman, S. (1995), Redundant sensor network for linear processes, *AIChE Journal*, v.41, pp. 2237-2249.

Almasy, G. A. (1990). Principles of dynamic balancing. *AIChE Journal, v.36, pp.1321-1330.*

Almasy, G. A., Sztano, T., (1975), Checking and correction of measurements on the basis of linear system model, *Problems of Control and Information Theory*, v.4, pp. 57–69.

Almasy, G.A., Mah, R.S.H., (1984), Estimation of measurement error variances from process data. *Ind. Engng. Chem. Process. Dev.*, v.23, pp. 779-784.

Arora, N., Biegler, L. T., (2001), Redescending estimators for Data Reconciliation and Parameter Estimation, *Computers and Chemical Engineering*, v.25, pp. 1585-1599.

Bagajewicz, M. J., Jiang, Q., (1997), Integral Approach To Plant Linear Dynamic Reconciliation, *AIChE Journal*, v. 43, pp. 2546-2558.

Bagajewicz, M. J., Jiang, Q., (1998), Gross error modeling and detection in plant linear dynamic reconciliation, *Computers and Chemical Engineering*, v.24, pp. 1789-1809.

Bagajewicz, M., Q. Jiang, and M. Sanchez, (2000), Removing and Assessing Uncertainties in Two Efficient Gross Error Collective Compensation Methods. *Chemical Engineering Communications*, 1563-5201, V. 178 (1), pp. 1 – 20

Bagajewicz, M. J., (2000), A Brief Review of Recent Developments in data Reconciliation and gross Error Detection/Estimation, *Latin American Applied Research.*, v. 30, pp. 335-342.

Bagajewicz, M. and D. Rollins.(2002), On the Consistency of the measurement and GLR tests

for Gross Error Detection. *Comp. & Chem. Eng.*

Bagajewicz, M. J., Cabrera, E.,(2003), Data reconciliation in gas pipeline systems, *Industrial and Engineering Chemistry Research*, v. 42, pp. 5596-5606.

Bagajewicz, M. J.; Nguyen D., (2007), Stochastic Based Accuracy of Data Reconciliation Estimators for linear Systems. *Computers and Chemical Engineering*, 32, 1257–1269.

Bascur, O. A., Linares, R., (2006), Grade recovery optimization using data unification and real time gross error detection, *Minerals Engineering*, v. 19 pp. 696–702.

Beck, J. B., Arnold, K. J., (1977), *Parameter Estimation in Engineering Science*, John Wiley & Sons, USA

Benqlilou C., (2004). *Data reconciliation as framework for chemical Processes Optimizaion and control*, Thesis, Universitat Politecnica de Catalunya

Biegler L. T., (1997). Response, *Computers Chem. Engng*. Vol. 21, No 9,p. 1071.

Brown R. G., Hwang, P. Y. C. (1997). *Introduction to random signals and applied Kalman Filtering.* John Wiley & Sons, Inc.

Butler, R. J., McDonald, J. B., Nelson, R. D., White, S. B., (1990), Robust and Partially Adaptive Estimation of Regression Models, *Rev. Econ. Stat.*, v. 72, pp. 321-333.

Cao, S., Rhinehart, R.R., 1995, An Efficient Method for On-Line Identification of Steady-State, *Journal of Process Control*, v.5, pp. 363-374.

Charpentier, V., Chang, L. J., Schwenzer, G. M., Bardin M. C., (1991), An Online Data reconciliation system for crude and vacuum units. *NPRA Computer Conference*, Houston, Texas.

Chen, J., Bandoni, A., Romagnoli, J. A., (1997), Robust estimation of measurement error variance/covariance from process sampling data. *Computers and Chemical Engineering*, v. 21, pp. 593-600.

Chen, J., Romagnoli, J. A. (1998). A strategy for simultaneous dynamic data reconciliation and outlier detection. *Computers and Chemical Engineering*, 22 (4/5), 559–562.

Chen, J., Bandoni, A., Romagnoli, J. A., (1998), Outlier Detection in Process Plant Data, *Computers and Chemical Engineering*, v.22, pp. 641-646.

Chen, J., Chen Z., Su, H, Chi, J., (2001), Two step method for Gross Error Detection in Process Data, *Proceedings of the American Control Conference*, Arlington, VA, pp. 2121-2126

Crowe, C. M., Garcia Campos, Y. A., & Hrymak, A. (1983). Reconciliation of process flow rates by matrix projection—part I: the linear case. *AIChE J.*, 29 (6), 881–888.

Crowe, C. M., (1988), Recursive Identification of Gross Errors in Linear Data Reconciliation, *AIChE Journal*, v. 34, pp. 541-550.

Crowe, C. M., (1989), Test of maximum power for detection of gross errors in process constraints, *AIChE Journal*, v. 35, pp. 869-872.

Crowe, C. M., (1992), The Maximum-Power Test for Gross Errors in the Original Constraints in Data Reconciliation, *Canadian Journal of Chemical Engineering*, v. 70, pp. 1030-1036.

Crowe, C. M. (1996). Data reconciliation—Progress and Challenges. *Journal of Process Control*, 6, 89–98.

Crowe, C. M., (1996b), Formulation of Linear Data Reconciliation Using Information Theory, *Chemical Engineering Science*, v. 51, pp. 3359-3366

Congli, M; Hongye S., Jian C.. (2006). An NT-MT Combined method for Gross Error Detection and Data Reconciliation. *Chinese J. Chem. Eng.*,14(5), 592–596.

Darouach, M., Ragot, J., Fayolle, J., Maquin, D., (1986), Validation des measures par équilibrage de bilans matière, *International Journal of Mineral Processing*, v.17, pp. 273-285.

Darouach, M., Ragot, R., Zasadzinski, M., Karzakala, G., (1989), Maximum likelihood estimator of measurement error variances in data reconciliation, *IFAC*, AIPAC Symposium, v.2, pp. 135-139.

Darouach, M., & Zasadzinski, M. (1991). Data reconciliation in generalized linear dynamic systems. *AIChE J.*, 37 (2), 193–201.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statistical Society Series B*, 39, 1.

Devanathan S., Varderman, S. B., Rollins D. K., SR., (2005), Likelihood and Bayesian Methods for accurate identification of measurement biases in pseudo-state state processes, *Trans IChemE*, Part A, pp. 1391-1398

Devanathan, S., Rollins, D. K, Vardeman, S. B., (2000), A new approach for improved identification of measurement bias, *Computers and Chemical Engineering*, v. 24, pp. 2755-2764.

Devanathan, S., Vardeman, S. B., Rollins, D. K (2004), Likelihood and Bayesian methods for accurate identification of measurement biases in pseudo steady-state processes, *Trans*

IChemE, Part A. Chemical Eng. Research and Design, 83(A12), pp. 1391-1398.

Dovi, V.G., Del Borghi, A., (2001), Rectification of Flow Measurements in Continuous Process Subject to Fluctuations, *Chemical Engineering Science*, v. 56, pp. 2851-2857.

Dunia, R., Qin, J. S., Edgar, T. F., McAvoy, T. J., (1996), Use of principal component analysis for sensor fault identification, *Computers and Chemical Engineering*, v. 20, pp. 713-718.

Fair, R. C., (1974), On the robust estimation of econometric models, *Annals of Economic and Social Measurement*, v. 3, pp. 667–677.

Farris, R. H., Law, V. J., (1979), An efficient computational technique for generalized application of maximum likelihood to improve correlation of experimental data, *Computers and Chemical Engineering*, v. 3, pp. 95–104.

Gauss, C. F., (1809), *Theoria Motus Corporum Coelestium*, Translation reprinted as Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections, 1963, New York, Dover Publications.

Gelb, A., (1969), *Appplied Optimal Estimation*, MIT Academic Press, Cambridge, MA.

Gnedenko, B, A. (2008), *Teoria da Probabilidade*, Tradução da série de textos clássicos da American mathematical society, 1° Edição em Português, Editora Ciência Moderna, Rio de Janeiro

Guoyong, L., & Chen, Z. (1985). Projection pursuit approach to robust dispersion matrices and principal components. *Journal of the American Statistical Association,* 80, 759.

Gupta, G., Narashiman, S., (1993), Application of Neural Networks for Gross Error Detection, *Industrial and Engineering Chemistry Research*, v. 32, pp. 1651-1667.

Hampel, F. R., (1971), A General Qualitative Definition of Robustness, *Annals of Mathematical Statistics*, v. 42, pp. 1887-1896.

Hampel, F. R., (1974), The influence Curve and its Role in Robust Estimation, *Journal of American Statistical Association*, v. 69, pp. 383-393.

Hampel, F. R., (1985), The Breakdown Points of the Mean Combined with some Rejection Rules, *Technometrics*, v.27, pp. 95-107.

Hampel, F.R., Ronchetti, E. M., Rousseau, P. J., Stahel, W. A., (1986), *Robust Statistics- The approach based on influence functions*. New York, Jonh Wiley.

Harikumar, P., Narashiman, S., (1993), A method to incorporate Bounds in Data Reconciliation and Gross Error Detection – II. Gross Error Detection Strategies, *Computers and Chemical Engineering*, v. 17, pp. 1121-1128.

Himmelblau, D. M., (1978), *Fault detection and diagnosis in chemical and petrochemical processes*, Elsevier Scientific Publishing Company, Amsterdam.

Hlavácek, V., (1977), Analysis of a Complex plant-Steady state and transient Behavior, *Computers and Chemical Engineering*, v.1, pp. 75-80.

Hsu, H., (1997), *Theory and problems of probability, random variables and random processes*. Schaum's outline series, McGrw-Hill, USA

Huber, P. J. (1964). A robust estimation of a location parameter. *Annals of Mathematics and Statistics*, 35, 1.

Huber, P. J. (1981). *Robust statistics*. New York, Wiley.

Huber, P. J. (1973), Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, v. 1, pp. 799-821.

Huber, P. J., (1964), Robust Estimation of a Location Parameter, *Annals of Mathematical Statistics*, v. 35, pp. 73-101.

Jazwinski A. H., (1970), *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Jiang, Q., Bagajewicz, M. J., (1999), On a Strategy with collective Compensation for multiple gross error estimation in Linear Steady-State Reconciliation. *Ind. Engng. Chem. Res.*, 38, 2119-2128.

Jiang, Q., Sánchez, M., Bagajewicz, M., (1999), On the Performance of Principal Component Analysis in Multiple Gross Error Identification, *Ind. Engng. Chem. Res.*, v.38, pp.2005-2012.

Johnson, L. P. M., Kramer, M. A., (1995), Maximum Likelihood Data Rectification: Steady-State Systems, *AIChE Journal*, v. 41, pp. 2415-2426.

Johnston, L. P. M., & Kramer, M. A. (1994). Probability density estimation using elliptical basis function. *AIChE J.*, 40, 1639.

Johnston, L. P. M., & Kramer, M. A. (1995). Maximum likelihood data reconciliation, steady state systems. *AIChE J*, 41, 2415.

Jordache, C., Mah, R. S. H., Tamhane, A. C., (1985), Performance studies of the measurement test for detection of gross errors in process data, *AIChE Journal*, v. 31, pp.1187-1201.

Kalman, R. E., (1960), A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering*, v. 82, pp. 35-45.

Kao, C. S., Tamahane, A. C., & Mah, R. S. H. (1992). Gross error detection in serially correlated process data. *Industrial Engineering and Chemical Research*, 31, 254–262.

Karjala, T. W., Himmelblau, D. M., (1994), Dynamic data rectification by recurrent neural networks versus traditional methods, *AIChE Journal*, v. 40, pp. 1865-1875.

Keller, J.Y., Darouach, M., Krzakala, G., (1994), Fault Detection of Multiple Biases or Process Leaks in Linear Steady State Systems, *Computers and Chemical Engineering*, v. 18, pp.1001-1004.

Kelly, J. D., (2004b), Techniques for solving industrial nonlinear data reconciliation problems, *Computers and Chemical Engineering*, v. 28, pp.2837-2843.

Kim I.-W., M. J. Liebman and T. F. Edgar, (1990). Robust error-in-variables estimation using nonlinear programming techniques. *AIChE J.* 36.985- 993.

Kim L-W., M. J. Liebman and T. F. Edgar, (1991). A sequential error-in-variables method for nonlinear dynamic systems. *Compupers and Chem. Engng* 15, 663-670.

Kim, I. W., Kang, M.S., Park, S., et al., (1997), Robust Data Reconciliation and Gross Error Detection: The Modified MIMT Using NLP, *Computers and Chemical Engineering*, v. 21, pp.775-782.

Kim, I. W., Liebman, M. J., Edgar, T. F., 1990, Robust Error-in-Variables Estimation Using Nonlinear Programming Techniques, *AIChE Journal*, v. 36, pp. 985-993.

Knepper, J. C., & Gorman, J. W. (1980). Statistical analysis of constrained data sets. *AIChE J.*, 37, 260.

Kolmogorov A. N., (1956), *Foundations of the theory of probability*, Chelsea Publishing Company, New York. 2nd English Edition

Kongsjahju, R., Rollins D. K., Bascuñana, M. B., (2000), Accurate Identification of biased measurements under serial correlation, *Trans IChemE*, vol. 78, Prt A, pp. 1010

Kramer, M. A. (1992). Autoassociative neural networks. *Computers and Chemical Engineering*, 16, 313.

Kretsovalis, A., Mah, R. S., (1987), Effect of Redundancy on Estimation Accuracy in Process Data reconciliation, *Chemical Engineering Science*, v.42, pp. 2115-2121.

Kretsovalis, A., Mah, R. S., (1988a), Observability and Redundancy Classification in Generalized Process Networks- I. Theorems, *Computers and Chemical Engineering*, v.12, pp. 671-687.

Kretsovalis, A., Mah, R. S., (1988b), Observability and Redundancy Classification in Generalized Process Networks- II. Algorithms, *Computers and Chemical Engineering*, v.12, pp. 689-703.

Kuehn, D. R., & Davidson, H. (1961). Computer control. II. Mathematics of control. *Chemical Engineering Progress*, 57, 44.

Lathi, B. P., (1998), *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA.

Legendre, A.M., (1805), *On the Method of Least Squares*, Translated from the French in D. E. SMITH, ed., A Source Book in Mathematics, 1959, pp. 576-79, New York, Dover Publications.

Lid T., (2007), *Data reconciliation and optimal operation with applications to refinery processes*, Thesis, Norwegian University of Science and Technology

Lid, T., Skogestad, S., (2008), Scaled steady State Models for Effective On-Line Applications, *Computers and Chemical Engineering*, v.32, pp. 990-999.

Liebman, M. J., Edgar, T. F., & Lasdon, L. S. (1992). Efficient data reconciliation and estimation for dynamic processes using nonlinear programing techniques. *Computers and Chemical Engineering*, 16 (10/11), 963–986.

Liebman, M. J., Edgar, T. F., (1988), Data reconciliation for Nonlinear Processes, *AIChE Annual Meeting*. Washington DC.

Liebman, M. J., Edgar, T. F., Lasdon, L. S., (1992), Efficient Data Reconciliation and Estimation for Dynamic Processes Using Nonlinear Programming Techniques, *Computers and Chemical Engineering*, v. 16, pp. 963-986.

Limqueco L. C., Kantor J. C. (1990). Nonlinear output feedback control of an exothermic reactor. *Computers and Chemical Engineering* 14, 427 .

Liptak, B. G., & Venczel, K. (1982). *Instrument engineers handbook*. Randor, PA: Chilton Book Company.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Madron, F., (1979), Material Balance Calculations of Fermentation Process, *Biotechnology and Bioengineering*, v. 21, pp. 1487-1490.

Madron, F., (1985), A New Approach to the Identification of Gross Error in Chemical Engineering Measurements, *Chemical Engineering Science*, v.40, pp. 1855-1860.

Madron, F., Veverka, V., (1992), Optimal Selection of Measuring Points in Complex Plants by Linear Models, *AIChE Journal*, v.38, pp. 227-236.

Mah, R. S. H. (1987). *Foundations of computer aided process operations*. New York, Springer.

Mah, R. S. H. (1990). *Chemical process structures and information flows.* London, Butterworths

Mah, R. S. H., & Tamhane, A. C. (1982). Detection of gross error in process data. *AIChE J.*, 28, 828.

Mah, R. S. H., (1990), *Chemical Process Structures and Information Flows*, 1 Ed. Stoneham. Butterworth.

Mah, R. S. H., (1997), Letter to the Editor, *Computers chem. Engng*. Vol. 21, No 9, p. 1069

Maquin, D., Adrot, O., Ragot, J., (2000), Data Reconciliation with Uncertain Models, *Isa Transactions*, v. 39, pp. 35-45.

Maronna R. A., Maritn R. D., Yohai, V. J., (2006), *Robust Statistics Theory and Methods*, John e Wiley Sons, England.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics,* 4, 51.

Martinez, A. Martinez W., (2002), *Computational Statistics Handbook with Matlab*, Chapman e Hall, Florida, USA

McBrayer, K. F., Edgar, T. F., (1995), Bias Detection And Estimation In Dynamic Data Reconciliation, *Journal of Process Control*. v. 5, pp. 285-289.

McBrayer, K., & Edgar, T. F. (1995). Bias detection and estimation in dynamic data reconciliation. *Journal of Process Control*, 5 (4), 285–289.

Mei, C., Su, H., Chu, J., (2006), An NT-MT Combined Method for Gross Error Detection and Data Reconciliation, *Chinese J. Chem. Eng.*, v. 14, pp. 592-596.

Mood, A. M., (1974), *Introduction to the theory of statistics*, McGraw-Hill

Morad K., Young B. R., Svrcek Q. Y.,(2005). Rectification of plant measurements using a statistical framework, *Computers and Chem.Engng*, 29, 919-940

Morad, K. (2000). *Probabilistic process data rectification*. Ph.D. thesis, University of Calgary.

Morad, K., Svrcek, W. Y., & McKay, I. (1999). A robust direct approach for calculating measurement error covariance matrix. *Computers and Chemical Engineering*, 23, 889.

Morad, K., Svrcek, W. Y., & McKay, I. (2000). Probability density estimation using incomplete data. *ISA Transactions*, 39, 379.

Narashiman S., Mah, R. S. H., (1989), Treatment of general steady-state process models in Gross error identification. *Computers Chem. Engng*, Vol. 13, N° 7, pp. 851-853

Narashiman, S., Jordache, C., (2000), *Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data*, Gulf Professional Publishing. Houston, TX,

Narasimhan, S., & Mah, R. S. H. (1987). Generalized likelihood ratio method for gross error identification. *AIChE J., 33 (9), 1514–1521*.

Narashiman, S., Shah, S., (2008), Model identification and error covariance matrix estimation from noisy data using PCA, *Control Engineering Practice*, v. 16, pp. 146-155.

Nogita, S. (1972). Statistical tests and adjustment of process data. *Industrial and Engineering Chemistry Process, Design and Development*, 11, 197.

Optimization Toolbox. The MathWorks Inc., 3, 4.2 version Apple Hill Drive,

Özyurt, D. B., Pike, R. W., (2004), Theory and practice of simultaneous data reconciliation and gross error detection for chemical process, *Computers and Chemical Engineering*, v. 28, pp. 381-402.

Pai, C. C. D., & Fisher, G. D. (1988). Application of the Broyden's method to reconciliation of non-linearly constrained data. *AIChE J.*, 34, 873.

Paulino C. D., Turkman, M. A. A., Murteira B., (2003), *Estatística Bayesiana*, Fundação Calouste Gulbenkian, Lisboa

Pearson, K., (1924), Historical Note on the Origin of the Normal Curve of Errors, *Biometrika*, v. 16, pp. 402-404.

Pearson, R. K., (2001), Exploring process data, *Journal of Process Control*, v. 11, pp. 179–194.

Priebe, C. E. (1994). Adaptive mixtures. *Journal of American Statistical Association*, 89, 796.

Priebe, C. E., & Marchette, D. J. (1991). Adaptive mixtures: Recursive non-parametric pattern recognition. *Pattern Recognition*, 24, 1197.

Priebe, C. E., & Marchette, D. J. (1993). Adaptive mixtures density estimation. *Pattern Recognition*, 26, 771.

Ragot, J., Chadli, M., Maquin, D., 2005, Mass Balance Equilibration: A Robust Approach Using Contaminated Distribution, *AIChE Journal*, v. 51, pp.1569-1575.

Ragot, J., Maquin D., Alhaj-Dibo M., (2005), Linear mass equilibration: a new approach for an old problem, *ISA transactions*, 23-24

Ramamurthi Y., Sistu P. B. and Bequette B. W., (1993). Control-Relevant dynamic data reconciliation and parameter estimation. *Computers and Chem. Engng*, Vol. 17, No. I, pp. 41-59.

Rao, C. V., Rawlings, J. B., (2002), Constrained Process Monitoring: Moving-Horizon Approach, *AIChE Journal*, v. 48, pp. 97-107.

Rao, R. R., & Narasimhan, S. (1996). Comparison of techniques for data reconciliation of multi component process. *Industrial Engineering and Chemical Research*, 35, 1362.

Ray W. H., (1982). New approaches to the dynamics of nonlinear systems with implications for process system design. In *Chemical Process Control 2* (Seborg D. E. and T. F. Edgar, Eds), 246. United Engineering Trustees, New York.

Reilly, P., Carpani, R., (1963), Application of statistical theory of adjustment to material balances, In *Proceedings of the 13th Canadian Chemical Engineering Conference*, Montreal, Quebec.

Renganathan, T., Narashiman, S., (1999), A Strategy for Detection of Gross Errors in Nonlinear Processes, *Industrial and Engineering Chemistry Research*, v. 38, pp. 2391-2399.

Ripps, D. L. (1965). Adjustment of experimental data. *Chemical Engineering Progress Symposium Series*, 61, 8.

Rollins , D. K., Cheng, Y., Devanathan, S., (1996), Intelligent Selection of hypothesis tests to enahnce gross error identification, *Computers chem. Engng*, Vol 20, N° 5, 217-529

Rollins, D. K., Davis, J. F., (1992), Unbiased Estimation of Gross Error in Process Measurements, *AIChE Journal*, v. 38 pp. 563-572.

Rollins, D. K., Davis, J. F., (1993), Gross error detection when variance-covariance matrices are unknown, *AIChE Journal*, v. 39, pp. 1335-1341.

Rollins, D. K., Devanathan, S., (1993), Unbiased Estimation in dynamic data reconciliation, *AIChE Journal*, v. 39, pp. 1330-1334.

Rollins, D. K., Devanathan, S., Bascuñana, M. V. B., (2002), Measurement bias detection in linear dynamic systems, *Computers and Chemical Engineering*. v.26, pp. 1201-1211.

Romagnoli, J. A., & Stephanopoulos, G. (1981). Rectification of process measurement data in the presence of gross errors. *Chemical Engineering Science*, 36 (11), 1849–1863.

Romagnoli, J. A., Sanchez, M. C., (2000), *Data Processing and Reconciliation for Chemical Process Operations*, Academic Press. San Diego.

Romagnoli, J. A., Stephanopoulos, G., (1980), On the Rectification of Measurement Errors for Complex Chemical Plants – Steady state Analysis, *Chemical Engineering Science*, v. 35, pp. 1067-1081.

Romagnoli, J. A., Stephanopoulos, G., (1981), Rectification of Process Measurement data in the Presence of Gross Errors, *Chemical Engineering Science*, v. 36, pp. 1849-1863.

Rosenberg, J., Mah, R. S. H, Jordache, C., (1986), Evaluation of schemes for detecting and identification of gross error in process data, *Industrial and Engineering Chemistry Research*., v. 26, pp. 3616-3631.

Rotman, J. J., *Advanced Modern Algebra*, (2002), Prentice Hall, USA

Safavi, A. A., Chen J., Tomagnoli, J. A., (1997),Wavelet-based Density Estimation and application to process Monitoring, *AIChe Journal*, 43, 1227

Salau, N. P., Secchi, A. R., Trierweiler, J. O.. (2007). Five formulations of extended kalman filter: Which is the best for D-RTO? *ESCAPE 17*

Sánchez, M., Bandoni, A., Romagnoli, J. A., (1992), PLADAT: A Package for Process Variable Classification and Plant Data Reconciliation, *Computers and Chemical Engineering*. vol. 16, pp. 499-506.

Sánchez, M., Romagnoli, J. A., (1996), Use of Orthogonal Transformations in Data Classification – Reconciliation, *Computers and Chemical Engineering*. v. 20, pp. 483-493.

Sánchez, M., Romagnoli, J. A., Jiang, Q., Bagajewicz, M., (1999), Simultaneous Estimation of Biases and Leaks In Process Plants, *Computers and Chemical Engineering*, vol. 23, pp. 841-857.

Scharaa, O. J., Crowe, C. M., (1998) The numerical solution of bilinear data reconciliation problems using unconstrained optimization methods, *Computers and Chemical Engineering*, v. 22, pp. 1215-1228.

Serth, R. W., & Heenan, W. A. (1986) Gross errors detection and data reconciliation in steam-metering systems, *AIChE Journal*, v. 32, pp. 733-742.

Serth, R. W., Valero C. M., Heenan, W. A. (1987), Detection of gross errors in nonlinearly constrained data: a case study, *Chemical Engineering Communications* v.51, pp. 89-104.

Shiryaev, A. N., (1980), *Probability - Graduate Texts in Mathematics*, 2nd Edition, Springer.

Signal Processing, The MathWorks Inc., 3, 4.2 version Apple Hill Drive

Stanley G. M. and R. S. H. Mah, (1981), Observability and redundancy in process data. *Chem. Egng Sci.* 36, 259-272

Statistical Toolbox, The MathWorks Inc., 3, 4.2 version Apple Hill Drive

Stigler, S. M., (1981), Gauss and the invention of least squares , The Annals of Statistics, v. 9, 465-474.

Tamhane, A. C., & Mah, R. S. H. (1985). Data reconciliation and gross error detection in chemical process network. *Technometrics*, 27, 409.

Tamhane, A. C., (1982), A Note on the use of Residuals for Detecting an Outlier in Linear Regression, Biometrika, v. 69, pp. 488-489.

Tamhane, A. C., Mah, R. S. H., (1985), Data Reconciliation and Gross error Detection in Chemical Process Networks, Technometrics, v. 27, pp. 409-422.

Terry, P.A., Himmelblau, D. M., (1993), Data Rectification and Gross Error Detection in a Steady-State Process via Artificial Neural Networks, Industrial and Engineering Chemistry Research, v. 32, pp. 3020-3028.

Tjoa, I. B., & Biegler, L. T. (1991). Simultaneous strategies for data reconciliation and gross error detection of non-linear systems. *Computers of Chem. Engng*, 15, 679.

Tjoa, I. B., & Biegler, L. T. (1992). Reduced successive quadratic programming strategy for errors in variables estimation. *Computers of Chem. Engng*, 16, 525.

Tjoa, I. B., Biegler, L. T., (1991a), Simultaneous Solution and Optimization Strategies for Parameter Estimation of Differential-Algebraic Equation Systems, *Industrial and Engineering Chemistry Research*, v. 30, pp. 376-385.

Tong, H., Crowe, C. M., (1995), Detection of Gross Errors in Data Reconciliation by Principal Component Analysis, *AIChE Journal*, v. 41, pp. 1712-1722.

Tong, H., Crowe, C. M., (1996), Detecting persistent gross errors by sequential analysis of principal components, *AIChE Journal*, v. 41, pp. 1712-1722.

Trierweiler, J.O., Farenzena, M., (2006), Uma visão geral das tecnologias atualmente aplicadas em controle avançado de processos industriais, *Revista de controle e instrumentação*

Upp, E. L.; LaNasa, P. J., (2002), *Fluid flow measurements*, 2nd edition, Boston: Gulf Professional Publishing

Václavek, V., 1969, Studies on System Engineering – III optimal Choice of The Balance Measurements in Complicated Chemical Engineering Systems, Chemical Engineering Science, v. 24, pp. 947-955.

Václavek, V., Loucka, M., (1976), Selection of Measurements necessary to Achieve Multicomponent Mass Balances in Chemical Plant, *Chemical Engineering Science*, v. 31, pp. 1109-1205.

Verveka, V., (1992), A method of reconciliation of measured data with nonlinear constraints, Applied Mathematics and Computation, v. 49, pp. 141-176.

Verveka, V.V., Madron, F., (1997), *Material and Energy Balancing in the Process Industries: From Microscopic Balances to Large Plant*, Elsevier Science, Amsterdam, The Netherlands.

Wang, D., Romagnoli, J. A., (2003), A framework for robust data reconciliation based on a generalized objective function, *Industrial and Engineering Chemistry Research*, v. 42, pp. 3075-3084.

Wang, F., Jia, X., Zheng, D., Yue, J., (2004), An improved MT-NT method for gross error detection and data reconciliation, *Computers and Chemical Engineering*, v. 28, pp. 2189-2192.

Wang, N. S., Stepahnopoulos, G. N., (1983), Application of Macroscopic Balances to the Identification of Gross Measurements Errors, *Biotechnology and Bioengineering*, v. 25, pp. 2177-2208.

Watanabe, K., & Himmelblau, D. M. (1982). Instrument fault detection in systems with uncertainties. *Int. J. Systems Sci.*, 13 (2), 137.

Weiss, G. H., Romagnoli, J. A., Islam, K. A., (1996), Data Reconciliation – An Industrial Case Study, *Computers and Chemical Engineering*, v. 20, pp.1441-1449.

Willsky, A. S. (1976). A survey of design methods for failure detection in dynamic systems. *Automatica*, 12, 601–611.

Willsky, A. S., & Jones, H. L. (1974). a generalized likelihood ratio approach to state estimation in linear systems subject to abrupt changes. *Proc. IEEE Conf. Decision and Control* (p. 846).

Yang, Y., Ten, R., Jao, L., (1995), A Study of Gross Error Detection and Data Reconciliation in Process Industries, *Computers and Chemical Engineering*, v. 19, pp. 217-222.

Zhang, P., Rong, G., Wang, Y., (2001), A new method of redundancy analysis in data reconciliation and its application, *Computers and Chemical Engineering*, v. 25, pp. 941-949.

Zhou, L., Su, H., Chu, J., (2006), A New Method to Solve Robust Data Reconciliation in Nonlinear Process, *Chinese Journal of Chemical Engineering*, v. 14, pp.357-363.