ANÁLISE TEÓRICA E PRÁTICA DE IMPLEMENTAÇÕES DE SISTEMAS DE OTIMIZAÇÃO EM TEMPO REAL (RTO)

André Domingues Quelhas

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

Orientador: José Carlos Costa da Silva Pinto

Rio de Janeiro

Dezembro de 2013

ANÁLISE TEÓRICA E PRÁTICA DE IMPLEMENTAÇÕES DE SISTEMAS DE OTIMIZAÇÃO EM TEMPO REAL (RTO)

André Domingues Quelhas

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

Examinada por:

_____
Prof. José Carlos Costa da Silva Pinto, D.Sc.

_____
Dr. Antonio Carlos Zanin, D.Sc.

_____
Prof. Darci Odloak, D.Sc.

_____
Prof. Evaristo Chalbaud Biscaia Junior, D.Sc.

_____
Prof. Príamo Albuquerque Melo Junior, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2013

Quelhas, André Domingues

Análise Teórica e Prática de Implementações de Sistemas de Otimização em Tempo Real (RTO) / André Domingues Quelhas. – Rio de Janeiro: UFRJ/COPPE, 2013

X, 265 p.; il.; 29,7 cm.

Orientador: José Carlos Costa da Silva Pinto

Tese (doutorado) – UFRJ/COPPE/ Programa de Engenharia Química, 2013.

Referências Bibliográficas: p. 239-247.

1. Otimização de Processos. 2. Estimação de Parâmetros. 3. Detecção de Estado Estacionário. I.Pinto, José Carlos Costa da Silva. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Química. III. Título.

AGRADECIMENTOS

Esta tese foi escrita e pensada ao longo de um período de acontecimentos muito significativos em minha vida pessoal. Representou a despedida de muitos modos de pensar antigos e exigiu diversas renúncias.

Meu esforço só frutificou em virtude de eu estar cercado de pessoas que ajudaram, com sua colaboração direta ou mesmo que só com sua presença, para que meu caminho estivesse sempre aplainado. Não posso deixar de citar minha família, minha esposa e meu filho Estêvão, que nasceu no mês em que este trabalho foi iniciado, e que tem acompanhado todos os meus pensamentos desde então.

Muito deste trabalho deve-se à presença motivadora do meu orientador, José Carlos, do qual só lamento não havê-lo conhecido antes em minha trajetória profissional. Sua capacidade de perceber o que cada um tem de melhor e fazê-lo seguir em frente foram aspectos sem os quais muito desta produção não teria existido.

A presença de um colega e amigo que possa entender suas dificuldades e colocar-se à disposição para auxiliar no desfecho dos problemas pode ser a diferença entre o sucesso ou o fracasso. Contei com a colaboração imprescindível do Fernando Esteves, que fez este papel com maestria e auxiliou a mostrar os caminhos do Linux e do Picloud, sem os quais o intensivo uso computacional e a paralelização na nuvem requeridos para os cálculos não teriam sido finalizados a tempo desta defesa ocorrer neste século ☺

Numquam periculum sine periculo vincitur
O perigo nunca é vencido sem perigo

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

ANÁLISE TEÓRICA E PRÁTICA DE IMPLEMENTAÇÕES DE SISTEMAS DE OTIMIZAÇÃO EM TEMPO REAL (RTO)

André Domingues Quelhas

Dezembro/2013

Orientador: José Carlos Costa da Silva Pinto

Programa: Engenharia Química

Este trabalho analisa a fundamentação teórica que ampara o funcionamento de sistemas comerciais de Otimização em Tempo Real (RTO) usados na indústria, corroborando esta análise com dados de implementações reais em refino de petróleo e com um estudo de caso. São mostradas evidências das vulnerabilidades matemáticas do método de otimização em duas etapas, assim como dos métodos de detecção de estacionariedade. É feita a proposição conceitual e prática de um teste de detecção de adequabilidade de sinais que substitui o conceito de detecção de estacionariedade. É mostrado que: 1) Mesmo para um sistema de RTO estacionário é imprescindível o uso de uma representação dinâmica para a correta verificação da adequabilidade dos sinais; 2) As vulnerabilidades de RTO em duas etapas não podem ser eliminadas, apenas mitigadas, e para isto sua estrutura deve ser convenientemente projetada, conforme mostrado neste texto; 3) Dada a variabilidade gerada pela violação dos requisitos de eficiência e consistência estatística e à natureza da operação em malha fechada, os parâmetros estimados pelos sistemas de RTO devem ser pensados e projetados como coadjuvantes úteis para o aumento da robustez do sistema, e não como entidades com significado físico que auxiliem no diagnóstico da planta; 4) Todo projeto deveria levar em conta o cômputo dos custos decorrentes das vulnerabilidades matemáticas do RTO em duas camadas antes de descartar soluções mais simples, como o uso de condições auto-otimizáveis ou a otimização *off-line*.

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

A THEORETICAL AND PRACTICAL ANALYSIS OF REAL TIME OPTIMIZATION (RTO) IMPLEMENTATIONS

André Domingues Quelhas

December/2013

Advisor: José Carlos Costa da Silva Pinto

Department: Chemical Engineering

This thesis analyzes the theoretical foundation behind commercial software of Real Time Optimization (RTO) used in the industry. This analysis is supported by data from actual implementations in petroleum refining and exemplified with a case study. Evidences of mathematical vulnerabilities of the two step approach (parameter estimation/otimization) are presented. It is also proved that common testes of stationarity are unable to fulfill the requirements of RTO operation. It is proposed, both conceptually and practically, a test for suitability of signals that replaces the concept of steady state detection. This thesis also shows that: 1) Even for a stationary RTO system it is essential to use a dynamic model for proper verification of the suitability of process variables data; 2) Vulnerabilities of the two step approach cannot be eliminated. Careful attention to the structure design of RTO, as shown in this text, is the only way to at least mitigate such flaws; 3) Given the variability derived from the closed-loop operation as well from the violation of the requirements of statistical efficiency and consistency, the parameters estimated by the RTO systems must be selected as useful handles to increase system robustness rather than as entities with physical meaning for the sake of plant diagnosis; 4) Every project should take into account the calculation of the costs of the mathematical vulnerabilities of the two step approach before discarding simpler solutions, such as either self-optimized operation or off -line optimization.

Sumário

Simbologia

f conjunto de equações que inter-relacionam os elementos de **ZZ**

fm versão de ***f*** disponível ao usuário

f_{atr} funções de atribuição: assinalam os elementos de **ZZ** oriundos do conjunto de informações *a priori*

f_{med} funções de medição: assinalam os elementos de **ZZ** cujos valores advém da observação direta

g conjunto de relações funcionais de desigualdade que inter-relacionam os elementos de **ZZ**

gm versão de ***g*** disponível ao usuário

L valor da função objetivo econômica

Lm valor da representação disponível da função objetivo econômica

ψ função densidade de probabilidade

Conjuntos

dO subconjunto de **ZZ** que contém taxas de variação

I subconjunto de **ZZ** concernente à representação estacionária

Iap informações do processo disponíveis *a priori* sobre o processo

II variáveis de **ZZ** não relacionadas às variáveis de estado nem às suas derivadas

In subconjunto de **ZZ** que contém as variáveis necessárias

Iod informações do processo obtidas por observação direta

O subconjunto de **ZZ** que contém as variáveis de estado

OO união de **O** e **dO**

Q_a informações aparentes acumuladas ao longo da evolução do processo

Q_a^+ representação observável possível da história do processo até o instante atual

Rto conjunto de escolhas referentes à estrutura do otimizador em tempo real

Z entidades matemáticas descritoras da representação estacionária do processo

Za informações adquiridas por observação direta

Zm entidades matemáticas descritoras da representação estacionária *disponível* do processo

ZZ entidades matemáticas descritoras do processo

θ modificadores não nulos de **Θ**, graus de liberdade do procedimento de adaptação do modelo

Θ modificadores aditivos ao conjunto **Z**

τ elementos algébricos de **II** irrelevantes para a representação estacionária

Índices de elementos em conjuntos

apf informações *a priori* falsas

apv informações *a priori* verdadeiras

atr variáveis de definidas pelas equações de atribuição

crp variáveis sujeitas à corrupção da informação

ds elementos de **ZZ** que representam taxas de variação

df graus de liberdade da otimização da função objetivo econômica

dual variáveis consequentes também obtidas por observação direta

ee variáveis selecionadas para comporem o teste de escionariedade

est subconjunto de **Est** sujeito à atualização

Est variáveis necessárias, modificadas ao longo do tempo, e que não são medidas

in variáveis necessárias para a descrição de **ZZ**

fix variáveis de **ZZ** constantes ao longo do tempo

ms variáveis obtidas por observação direta

ms^{-} subconjunto de **ms** que é efetivamente usado para atribuir valores a **ZZ** por meio das funções de medição

ocu subconjunto de **in** que é variável ao longo do tempo e inacessível à observação direta

out variáveis consequentes de **ZZ**

rec subconjunto de **Rec** sujeito à atualização

Rec variáveis observadas diretamente e sujeitas à corrupção

s variáveis de estado

var variáveis de **ZZ** ou de **Z** que cujos valores sofrem mudanças em relação ao início da operação

upd subconjunto de **Upd** sujeito à atualização

Upd união de **Rec** e **Est**

1. Introdução

Os processos químicos industriais são formados por complexos arranjos de uma miríade de equipamentos e tubulações, onde diversas transformações físicas e químicas resultam em produtos destinados a atender a demanda de uso sob condições financeiras vantajosas. Devido a esta complexidade inerente, o simples ato de vender um produto expressa a materialização de um enorme conjunto de decisões. A chave para o sucesso consiste na coordenação cuidadosa de todas as alternativas possíveis, focando o máximo de desempenho em termos de lucro, segurança e confiabilidade.

A natureza assíncrona do processo de tomada de decisões torna a coordenação das ações muito difícil. Isto ocorre pois as decisões operacionais são subsidiadas por informações produzidas por fontes que operam em fequencias distintas (instrumentação, laboratório, preços, suprimento, demandas) que alimentam sistemas que implementam decisões focadas em horizontes de tempo largamente variáveis (controles regulatório e avançado, otimizadores, planejamento, *scheduling*). Além desta falta de sincronia encontrada na operação cotidiana, a coordenação dos esforços ainda é fortemente condicionada por um grande conjunto de decisões irreversíveis, a maioria deles relacionado com *hardware*, feitas antes mesmo da partida da planta, incluindo a geometria de equipamentos e tubulações, seleção de materiais e assim por diante. Estas decisões preliminares moldam o grau de liberdade disponível para as rotinas de operação e otimização da planta e podem seriamente limitar as possibilidades de desempenho.

Na prática industrial comum, os sistemas automatizados (principalmente os protocolos de controle de regulação) são responsáveis pela maior parte das decisões de rotina de curto prazo relacionadas com a rejeição de perturbações e de rastreamento do ponto de ajuste. Se alguns graus de liberdade não são utilizados pelos níveis inferiores de automação, é possível empregá-los em outra camada de automação, a fim de perseguir ativamente o melhor desempenho de lucro ao longo do tempo de operação.

Por outro lado, é muito menos comum encontrar sistemas de otimização em tempo real (RTO) na indústria. Isto justifica-se, em parte, pelo fato de que a aplicação de procedimentos de RTO pode não ser apropriada para todos os processos [1]. Além disso, embora a idéia subjacente ao RTO seja fácil de entender e aceitar (a operação do processo deve ser otimizada em tempo real à medida que as condições de contorno e

parâmetros relevantes se modiquem), sistemas de RTO não são totalmente aceitos na indústria [2]. Isto deve-se ao fato de muitas implementações reais acabarem por mostrar-se "intensivas no uso de mão de obra, difíceis de implementar e serem descontinuadas facilmente" [3].

Na verdade, o termo *otimização em tempo real* constitui um conceito um tanto vago. No presente trabalho, ele será definido como a aplicação automatizada de decisões orientadas ao lucro operacional, com base em modelos rigorosos de processos não-lineares, e que são implementadas com uma frequência maior, em média, do que a ocorrência de distúrbios que conduzem o processo para o desempenho abaixo do ideal. Em processos químicos, o RTO é responsável por traduzir uma receita do produto da camada de *scheduling* para o melhor conjunto de valores de referência para o controle preditivo (MPC) camada de modelo.

A existência do RTO se justifica se existirem mudanças recorrentes nos cenários de operação que induzam a planta a condições menos favoráveis de desempenho econômico. Na indústria química, estas mudanças são comumente associadas à variabilidade dos preços de matérias-primas e produtos, sujeitos às condições de mercado e à logística de transporte; à mudança de qualidade de matéria-prima (condição típica de beneficiadores primários de recursos minerais); à mudança de qualidade dos produtos para o atendimento a diferentes mercados; à alteração do desempenho de equipamentos e processos e à mudança de condições ambientais.

1.1. Arquitetura do Otimizador em Tempo Real

A otimização *on-line* é incorporada ao controle regulatório de processos e ao controle avançado como um laço que fecha uma malha de controle mais exterior, como observado na Figura 1, extraída de [4]. O otimizador recebe premissas do sistema de programação da produção ou do *scheduling*, tais como: demandas de quantidade e qualidade dos produtos; preços de matéria-prima, produtos e utilidades industriais; restrições de suprimentos de matérias-primas, volume disponível para armazenamento de produtos etc.. Tendo como base estas informações, o laço de otimização é responsável pela proposição dos valores de referência para as camadas inferiores de controle.

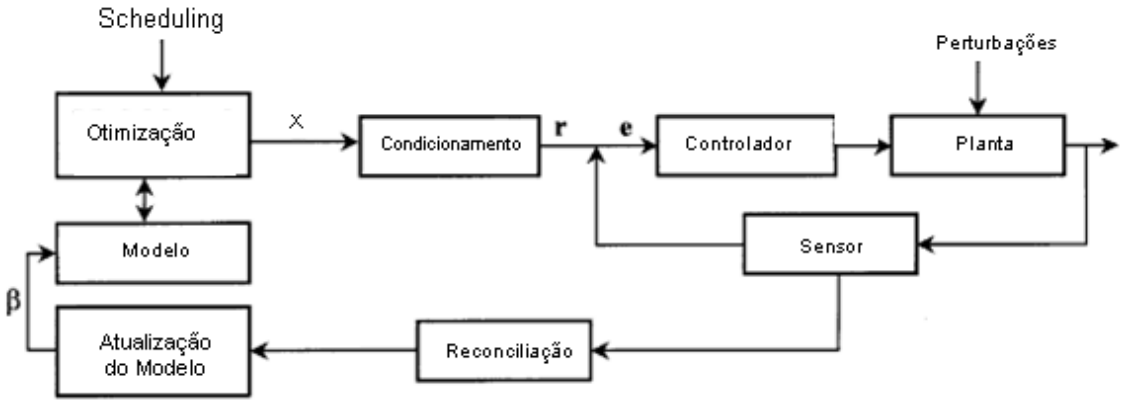


Figura 1 - Representação da atuação da otimização em tempo real integrada com a camada de controle de nível inferior.Variáveis de decisão: x; valores implementados: **r**; *set-points* dos controladores: **e**; parâmetros ajustados: **β**.

O vetor de valores de referência transmitido às camadas mais baixas da hierarquia de controle é escolhido em função do atendimento à condição extrema (máximo ou mínimo, a depender da conveniência) da função objetivo da etapa de otimização. Esta função objetivo é, invariavelmente, uma relação que vincula as condições operacionais da planta à realidade financeira da indústria na qual a planta está inserida. Esta relação quantifica, em unidade monetária, o valor associado ao uso de determinado conjunto de graus de liberdade da operação da planta.

A freqüência de operação do RTO é, no máximo, igual a das camadas por ela comandadas, a depender da complexidade do modelo que representa o processo e da probabilidade de ocorrência de perturbações.

As ações implementadas durante um ciclo de execução do RTO estacionário, visto na Figura 2, podem ser assim resumidas:

- Detecção do estado estacionário: tendo como base um subconjunto das informações medidas do processo, decide-se se a variabilidade observada não compromete a premissa de estacionariedade requerida para assegurar a validade dos modelos de comportamento do processo;

- Reconciliação de dados/atualização do modelo: procedimento que retifica um subconjunto das informações medidas à luz da coerência com o modelo do processo e adapta um subconjunto dos seus parâmetros de modo a capturar mudanças das condições de operação;

- Otimização: previsão de qual conjunto de valores dos graus de liberdade disponíveis garante a operação sob máximo desempenho econômico;

- Condicionamento: análise crítica dos valores propostos pelo otimizador, de modo a verificar se a melhoria do desempenho prevista é significativa, e verificação de estacionariedade, antes da implementação dos valores otimizados das variáveis de decisão.

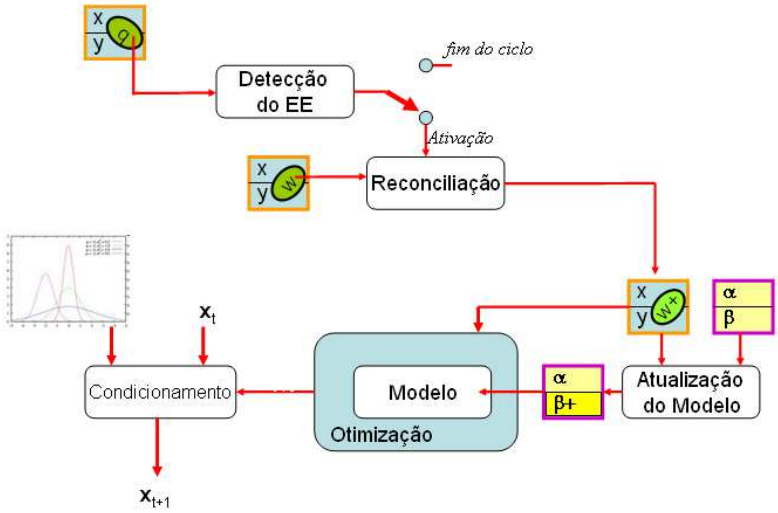


Figura 2 – Arquitetura típica de um otimizador estacionário em tempo real. Variáveis de entrada: **x**; de saída: **y**; valores reconciliados: **w**; parmetros estimados: **β**.

As operações executadas no núcleo do RTO consistem na adaptação do modelo de processo à evolução das condições de operação da planta e na proposição do conjunto de valores das variáveis de decisão que correspondem ao ponto extremo de uma métrica que seja mais conveniente no problema em questão.

1.2. Uso atual do RTO na Indústria

Considerando-se a operação de uma planta química, o modo como alterações nas condições de operação, ditas variáveis de entrada, $\mathcal{E}$, influenciam as variáveis de resposta, $\mathcal{R}$, é regido por um conjunto de equações fundamentais de grandezas conservativas, $\mathbf{f}(\mathcal{E},\mathcal{R},\mathcal{R}',\boldsymbol{\theta}) = \mathbf{0}$, cujas estrutura e respectivos parâmetros, $\boldsymbol{\theta}$, constituem o modelo de comportamento do processo e $\mathcal{R}'$ representa o vetor das taxas de variação $d\mathcal{R}/dt$. O problema de operação ótima da planta consiste em maximizar uma função de

desempenho, $L(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\gamma})$, que seja útil ao gerenciamento da planta, através da conveniente escolha do vetor de variáveis de estímulo **u**. Desta forma, o problema pode ser colocado sob a seguinte formulação matemática:

$$\begin{aligned} &\max_{u} L(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\gamma}) \\ &s.a.\ \mathbf{f}(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{R}}', \boldsymbol{\theta}) = \mathbf{0} \end{aligned} \tag{1}$$

A Equação (1) representa a essência do conceito de operação ótima, porém está associada a um cenário idealizado quanto à disponibilidade de informações do processo e quanto à extensão da liberdade de interferir nos estímulos que condicionam o valor ótimo de *L*, cujos parâmetros financeiros são expressos por $\boldsymbol{\gamma}$.

Em plantas industriais, a função de desempenho *L*, na maior parte das vezes, tem como núcleo o lucro financeiro, ou seja, a diferença entre receitas e despesas operacionais derivadas de cada escolha de **u**.

O modo mais simples, embora operacionalmente mais trabalhoso, de formular os valores de **u** que conduzem ao máximo valor de *L* se dá através de métodos de experimentação para atingir o ponto ótimo operacional, surgido a partir das propostas de Box [5,6] e que evoluíram para o método de busca direta [7,8,9]. Este método prevê a realização de vários experimentos na planta, modificando-se os valores de **u**, de modo a que se identifique o conjunto que conduza à melhor direção do gradiente da função de desempenho. Contudo, o elevado número de passos para alcançar o ótimo compromete o ganho econômico [10]. Exceto para plantas com poucos graus de liberdade e rápida resposta, há pouco apelo para o uso contemporâneo deste método.

Embora o problema proposto na Equação (1) se refira à otimização da trajetória do sistema no tempo, os RTOs comerciais não seguem esta abordagem, focando em modelos que descrevam o comportamento estacionário do processo. Este procedimento busca fugir das dificuldades associadas à obtenção de um modelo dinâmico fidedigno cujo controlador seja facilmente sintonizável [11].

Os diferentes tipos de otimizadores em tempo real propostos na literatura ao longo das últimas décadas foram classificados por Zhang e Forbes [12] e, mais recentemente, por Chachuat *et alii* [13,14]. A seguinte divisão tem sido proposta:

- algoritmos que adaptam o RTO à realidade da planta via modificação dos parâmetros do modelo com base nas medidas oriundas da instrumentação, representado pela otimização em duas etapas [15];

- algoritmos que, apesar de fazerem uso de medidas oriundas da instrumentação, não as usam para alterar diretamente os parâmetros do modelo, mas sim modificadores aditivos à função objetivo e restrições do problema de otimização, com vistas à garantir as condições de otimalidade [16,17,18,19];

- algoritmos que transformam o problema em algo similar a um problema de controle por retro-alimentação. Neste caso, não há sucessivas otimizações do modelo a cada intervenção, mas a manipulação de variáveis (ou combinações de variáveis) que mantenham o valor de determinada função *f* em dado valor de *set-point*. f depende de variáveis medidas e deve ser projetada de forma que sua manutenção no valor de referência esteja correlacionada ao ótimo de operação da planta [20,21,22].

O esquema mais comum (senão o único) de RTO na indústria é o da abordagem em duas etapas [15] usado nos principais *softwares* comerciais de uso industrial típico (Romeo 5.3 Invensys, Houston, TX, Aspenplus 7.1 Aspentech, Burlington, MA). Por esta razão, a maior parte da literatura técnica RTO é de alguma forma relacionada com este tema [23]. Os demais esquemas impõem uma carga de condições difícil de ser atendida em aplicações reais. Como exemplo, o método de transformação do problema de otimização em um de controle demanda a manutenção do conjunto de restrições ativas ao longo da operação, além de se basear na difícil elaboração de uma função 'otimizadora'. Os métodos baseados na adaptação de modificadores, apesar de basearem-se em uma fundamentação matemática mais sólida, que garante a optimalidade das soluções, exigem uma série de demoradas medições experimentais, a fim de avaliar gradientes de um grande conjunto de funções e variáveis. Dado o impacto considerável na produtividade, essas implementações são praticamente ausentes na prática industrial atual.

A técnica do RTO em duas etapas deve sua popularidade à idéia intuitiva que o suporta. Na primeira camada de otimização, a informação obtida das medições das variáveis de processo é usada para atualizar os parâmetros do modelo com base no melhor ajuste entre medidas e predições do modelo. Em seguida, a camada de

otimização gera um conjunto de valores das variáveis de decisão que são assumidas para conduzir processo para o seu melhor desempenho econômico.

1.3. RTO na presença de Incertezas

As incertezas de natureza aleatória contidas nas medições são transferidas aos resultados de cada camada de otimização. Assim sendo, todas as informações produzidas pelo sistema de RTO possuem natureza estocástica. Reconhecer este fato torna lícito questionar a validade estatística da proposição $\hat{\mathbf{u}}_k$ feita pelo otimizador durante a k-ésima intervenção no processo. Embora ela possa ser numericamente diferente da implementação anterior, $\hat{\mathbf{u}}_{k-1}^{imp}$, ambas podem ser realizações diferentes de um mesmo evento estocástico. Além disto, ainda que haja garantia estatística de que a nova implementação é distinta da anterior, isto pode não estar associado à expectativa de que o lucro L_{k+1} seja significativamente maior que L_k. Contudo, este tipo de preocupação tem sido pouco freqüente na literatura técnica e inexistente, de modo sistematizado, em implementações comerciais que sejam do conhecimento do presente autor.

Miletic e Marlin [24,25] abordam esta questão para problemas de otimização em tempo real baseados no método em duas etapas sujeitos apenas a restrições de igualdade. Eles sugerem o uso de um procedimento baseado em técnicas de controle estatístico de processos (SPC). O método foi posteriormente expandido de modo a dar conta do possível mau condicionament da matriz de covariâncias dos parâmetros estimados [26] e para levar em conta restrições de desigualdade no problema de otimização [27]. Nesse método, a especificidade do processo estudado está toda contida na matriz de covariâncias das estimativas das variáveis de decisão, que é resultado da propagação das incertezas de medição contidas na matriz de covariâncias das medições ao longo das camadas de estimação de parâmetros e de otimização. Deve-se notar que há diversas hipóteses importantes assumidas na confecção deste método, e que moldam o tipo de problemas ao qual se aplica. A saber, que os erros das medições possuem distribuição normal multivariável e que as transformações matemáticas que expressem a propagação dos erros também sejam lineares, de modo que as estimativas da variabilidade das variáveis de decisão também possuam distribuição normal multivariável.

Apesar da preocupação em condicionar o vetor de novos set-points sugeridos por cada iteração do RTO, a abordagem de Miletic e Marlin [24,25] consiste em um

procedimento reativo de análise de resultados, que não incorpora ao otimizador a capacidade de lidar com as incertezas das medições. Além disso, outros perigos ameaçam a qualidade dos resultados, que não só a desnecessária implementação sucessiva de set-points estatisticamente equivalentes. As soluções previstas pelo otimizador costumam situar-se próximas ou sobre as restrições $\mathbf{g}(\mathbf{u},\boldsymbol{\theta})\leq\mathbf{0}$, e isto é tão mais verdadeiro quanto mais linear for a estrutura do problema de otimização. Dadas as incertezas nas informações medidas e os erros de modelagem, o risco de que as soluções previstas sejam inviáveis na prática aumenta na medida que mais restrições se tornem ativas, isto é, $g_i(\mathbf{u},\boldsymbol{\theta})=0$, onde o subscrito indica cada uma das inequações em $\mathbf{g}$.

Levando esta questão em consideração, Loeblein e Perkins [28] sugerem o uso de um vetor cujos valores são respectivamente aditivos a cada uma das restrições com o fito de garantir um afastamento dos limites previstos. Deste modo, o vetor de afastamento $\boldsymbol{\beta}$ modificaria o conjunto de restrições de modo que:

$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) + \boldsymbol{\beta} \leq \mathbf{0} \tag{2}$$

Para o cálculo de $\boldsymbol{\beta}$ consideram-se as incertezas de erros de medição e variações aleatórias e determinísticas dos parâmetros. Uma vez caracterizadas estas variações, os elementos do vetor de afastamento podem ser calculados [28] a partir da otimização do modelo linearizado de modo a assegurar, para dado nível de confiança, que os limites das restrições não serão violados.

Apesar de o vetor de afastamento ser um mecanismo ativo de garantia de resultados ao incorporar salvaguardas no procedimento de otimização e não apenas apontá-los *a posteriori*, o seu cálculo é feito *offline*, supondo a manutenção das restrições ativas ao longo dos demais ciclos. Uma possível melhoria consiste em incorporar mais profundamente a natureza estocástica das informações medidas no processo de otimização, prescindindo de um vetor de afastamento das restrições. Recordando o problema da camada superior de otimização no formato determinístico:

$$\begin{aligned} \hat{\mathbf{u}} &= \max_{\mathbf{u}} \; L(\mathbf{u},\boldsymbol{\gamma}) \\ s.a.\; & \mathbf{g}(\mathbf{u},\boldsymbol{\theta}) \leq \mathbf{0} \end{aligned} \tag{3}$$

Deve-se considerar, porém, que o vetor das variáveis de decisão **u** e o vetor dos parâmetros **θ** são compostos, ao menos parcialmente, de grandezas que possuem natureza aleatória, fruto do processo de medição ou de reconciliação. Estes processos produzem valores que são, em verdade, estimativas das médias das variáveis. Desta forma, o sistema de inequações que compõe as restrições do problema corresponde, no processo tradicional de otimização, a operações aplicadas sobre a média estimada das informações: $\mathbf{g}(\hat{\mu}(\mathbf{u}), \hat{\mu}(\boldsymbol{\theta})) \leq \mathbf{0}$.

Este fato mostra que, sob a aparente simplicidade da formulação determinística da Equação (3), estão ocultos potenciais impactos estatísticos nos resultados da otimização. Se as funções $\mathbf{g}(\bullet,\bullet)$ consistirem apenas de transformações lineares, então a formulação das restrições propostas pela Equação (3) é equivalente a $E[\mathbf{g}(\mathbf{u},\boldsymbol{\theta})] \leq \mathbf{0}$. Isto significa que há garantia apenas de que cada restrição seja atendida em metade dos casos, o que está longe de ser satisfatório. Além disto, não há informação quanto ao grau esperado de afastamento das violações se considerada toda a faixa de distribuição de valores passível de ser obtida, que pode ser de grande monta. Por último, caso $\mathbf{g}(\bullet,\bullet)$ represente, parcial ou totalmente, um conjunto genérico de funções não lineares, não há formulação analítica prévia que possa indicar as conseqüências da otimização determinística realizada como em (3), o que pode esconder problemas ainda mais preocupantes nos resultados da otimização.

Com base no exposto, a formulação das restrições deveria levar em conta, explicitamente, a probabilidade de violação das restrições sob dado nível de confiança. Com isto em mente, as restrições definidas na Equação (3) deveriam ser expressas como:

$$P\left(\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) \leq \mathbf{0}\right) \geq \alpha \qquad (4)$$

A formulação das restrições ao problema de otimização expressa na Equação (4) assegura que os valores buscados pela otimização para as variáveis de decisão sejam capazes de garantir, no nível de confiança α, que *todas* as restrições sejam simultaneamente respeitadas. Esta formulação é conhecida como restrição de probabilidade conjunta, RPC (JPC em inglês). Embora esta formulação seja a mais rigorosa, as dificuldades de resolução por ela impostas fazem com que o problema mais relaxado, o da restrição de probabilidade individual, RPI (IPC em inglês) seja mais abordado na literatura [29], conforme a formulação (5). Neste caso, procura-se garantir a

probabilidade de cada restrição ser atingida isoladamente, mantendo a independência entre a probabilidade de cada restrição ser atingida.

$$P(g_i(\mathbf{u},\boldsymbol{\theta}) \le 0) \ge \alpha_i, \qquad i=1..ng \tag{5}$$

onde ng = dim(**g**) é o número de restrições

Outra mudança de formulação necessária diz respeito ao significado da operação de otimização da função objetivo. O significado de $\max(\mathrm{L}(\mathbf{u},\boldsymbol{\gamma}))$ fica pouco claro na medida em que $\mathrm{L}(\mathbf{u},\boldsymbol{\gamma})$ é uma variável aleatória, com respectiva função densidade de probabilidade. O modo mais comum de se encontrar um equivalente determinístico para a operação de otimização é o emprego de uma combinação linear da esperança matemática e da variância da função objetivo, como descrito em Darlington *et al.* [30]. A nova formulação do problema de otimização da Equação (3) que leva em conta a natureza aleatória das variáveis assume um formato de otimização conhecido como programação estocástica [31] e é apresentado sob a forma:

$$\begin{aligned}
&\hat{\mathbf{u}} = \min_{\mathbf{u}} \left\{(1-a)E\left[-L(\mathbf{u},\boldsymbol{\gamma})\right] + aV\left[L(\mathbf{u},\boldsymbol{\gamma})\right]\right\} \\
&s.a. \\
&P\left(\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) \le \mathbf{0}\right) \ge \alpha \\
&ou \\
&P(g_i(\mathbf{u},\boldsymbol{\theta}) \le 0) \ge \alpha_i, \qquad i=1..ng
\end{aligned} \tag{6}$$

onde $0 \le a \le 1$

Por simplificação e crença que o termo da variância não muda significativamente no ponto otimizado, ele é comumente omitido [29,32,33]. Por outro lado, o equivalente determinístico da função objetivo, composto como uma ponderação entre valor esperado da função objetivo (parcela da esperança matemática) e da robustez (parcela da variância) é questionado [34] devido ao fato de a variância não ser uma boa medida de risco por não levar em conta assimetrias da função densidade de probabilidade. Além disto, deve-se considerar que, em problemas como definidos pela Equação (6), a

variabilidade que torna o lucro maior é penalizada da mesma forma daquela que torna o lucro menor.

Embora seja o modo que responde de forma mais completa às questões inseridas em um problema de otimização sob a presença de incertezas, devido à sua complexidade de implementação e resolução computacional, os problemas de otimização estocástica não são encontrados em qualquer uso industrial até a presente data. Mesmo na literatura, os estudos de casos apresentados [29,35,36] lidam com funções de restrição nas quais a incerteza aparece linearmente, à exceção do trabalho de Kookos [33].

1.4. Configuração de sistemas de RTO – Estado da arte

O método de estimação em duas etapas não garante que o ponto ótimo será atingido nem que suas previsões serão estabilizadas após uma única execução do RTO mesmo na ausência de quaisquer perturbações ao processo, como será mostrado na Seção 3.1.1. Apesar disto, implementações comerciais são todas baseadas em métodos em duas etapas e sem informações mais elaboradas da caracterização das incertezas das medidas experimentais. Métodos como busca direta, ISOPE, e adaptação de modificadores [14] exigem um nível de detalhamento acerca das derivadas das variáveis na planta que é inviável de ser atingido.

O modo de funcionamento dos RTOs comerciais resume uma convicção muito comum na prática da engenharia: a de delegar diversas decisões a respeito da estrutura de funcionamento do RTO ao engenheiro de desenvolvimento, ao engenheiro de acompanhamento e à operação, na crença de que o conhecimento prático por eles acumulado os credencie a tomar as melhores decisões.

As primeiras decisões desta classe são aquelas tomadas ainda na fase de projeto da planta e relacionadas ao conjunto de variáveis medidas. Na maior parte das vezes, contudo, os projetos prevêem uma malha de instrumentação e sensoreamento que não visa a atender os requisitos vinculados ao bom funcionamento do RTO, no sentido de garantir que as variáveis estimadas o sejam com menor incerteza, que haja mais informação redundante, e que a estimativa do lucro seja mais fidedigna. Comumente, a malha de sensores segue o formato já utilizado em projetos similares, buscando fechar o balanço de massa e, quando muito, auxiliar em alguns controladores regulatórios. Na literatura, decisões deste tipo também não são comuns, e são baseadas em modelos

linearizados, freqüentemente associados à hipótese de constância do conjunto de restrições ativas [37,38,39].

Outra decisão crucial delegada ao usuário é a definição do conjunto de variáveis medidas que serão reconciliadas, assim como do conjunto dos parâmetros que serão estimados a cada novo ciclo do RTO. Quanto à questão das variáveis a serem reconciliadas, esta decisão costuma ser tomada com um grau de confiança incompatível com a parca quantidade e baixa qualidade das informações que subsidiam a resposta. Dado o elevado número de sensores em uma planta industrial, é muito improvável que a análise empírica do desempenho atual conduza ao conjunto ótimo de transdutores que realmente precisam ser reconciliados.

Estes critérios costumam basear-se: 1) na lembrança do histórico de manutenções *corretivas* (que, no mais das vezes, não são feitas devido à ocorrência de desvios sistemáticos, mas por falhas catastróficas); 2) na variabilidade das medidas (que pode estar relacionada ao filtro do sinal ou mesmo à variabilidade intrínseca do processo, que não fazem jus à "falta de confiança" nas medições); 3) baseado em idiossincrasias pessoais (cuja variabilidade é ainda maior que as das próprias medidas...), fruto do modelo de interpretação do comportamento da planta criado por cada um, devendo-se ressaltar que nem sempre o objetivo do processo e os padrões de operação são completamente entendidos, principalmente em sistemas multivariáveis.

A crítica referente à reconciliação de dados é válida também para a escolha dos parâmetros estimados, com algumas nuances adicionais. Diferentemente do problema de reconciliação, não há sensores cujas informações possam ser pretensamente escalonadas em diferentes níveis de acurácia. A decisão é tomada baseada na expectativa de variação dos parâmetros, e tal expectativa costuma ser criada a partir do suposto significado físico a eles vinculado. Nesta representação reduzida do problema, há espaço para a potencial contribuição da experiência do profissional envolvido na medida em que a suposta relação entre evidências físicas e as entidades matemáticas representadas pelos parâmetros permite a tradução de evidências empíricas em análogos abstratos que conduzam a adaptação do modelo à realidade.

Como exemplo desta correspondência entre as duas realidades, pode-se imaginar que alguém constate, ao fim de cada parada da unidade, o acúmulo de incrustações em determinado trocador de calor. Este fato faz com que se leve seriamente em consideração a necessidade de incluir o parâmetro associado à troca térmica no conjunto

daqueles que serão atualizados. A conclusão parece ser direta, porém há alguns problemas da vida real que limitarão a eficiência deste procedimento.

Em primeiro lugar, a incrustação no trocador não credencia este equipamento a ter um parâmetro selecionado entre aqueles que serão atualizados pois: a troca térmica neste equipamento pode afetar pouco a escolha dos *set-points* que levarão o processo ao lucro ótimo; a incerteza da estimativa deste parâmetro pode ser muito elevada; as condições de operação podem tornar pouco importantes a ocorrência de incrustação (em caso de elevada condutividade térmica do depósito e/ou elevados valores de tensão de cisalhamento).

É muito comum que os parâmetros sejam escolhidos para atualização devido ao seu pretenso significado físico, o que pode ajudar a alcançar dois objetivos de uma só vez: adaptar o RTO, tornando-o potencialmente mais capaz de perseguir o valor ótimo e também oferecer informações de diagnóstico sobre a operação dos equipamentos. Este desejo de usar os parâmetros para monitoração muitas vezes está ancorado em modelos que ignoram a multiplicidade de causas que podem originar as mesmas conseqüências.

Caso um modelo de troca térmica do tipo $\dot{m}C_p\Delta T = UA\Delta T_{\ln}$ tenha o parâmetro U atualizado, este coeficiente, mesmo que estimado com perfeição, retratará o efeito combinado de diversas outras influências, como mudanças na composição química da vazão, na condutividade térmica e na espessura da incrustação, assim como na tensão de cisalhamento. Acompanhar a evolução temporal das estimativas de U pode não auxiliar no diagnóstico de incrustação.

Por último fica a questão de como vincular corretamente a escolha do parâmetro associado às evidências físicas acumuladas pela experiência. Pode-se considerar um caso em que o modelo de troca térmica seja minucioso e leve em conta diversas medições **x** (vazões, temperaturas e pressões na entrada e saída do trocador) e parâmetros **p** (diâmetro de tubulações, área de troca, tensão de cisalhamento, resistência condutiva e convectiva, etc..), em uma estrutura funcional do tipo $\dot{Q} = f(\mathbf{x}, \mathbf{p})$. Motivado pela evidência física de incrustação, assume-se que há necessidade de adaptar esta equação às condições operacionais. Supondo-se que haja liberdade para atualizar mais de um parâmetro, quantos e quais serão escolhidos? A atualização do diâmetro interno da tubulação parece intuitiva, pois que baseada nas evidências físicas observadas. Porém, em vista do abordado no caso anterior, atualizar apenas este parâmetro pode ser contraproducente. Por outro lado, caso não haja variabilidade suficiente na informação,

decidir estimar muitos parâmetros resultará em estimativas de baixa qualidade. Em vista disto, como determinar o tamanho do vetor de parâmetros atualizáveis e como compor seus elementos? A experiência motivada pela evidência física não apresenta soluções a estas questões. Na verdade, nem mesmo as enuncia.

Estas considerações procuram evidenciar que, ainda que o conhecimento adquirido através do trabalho cotidiano seja de grande relevância e traga consideráveis ganhos à qualidade da atuação do engenheiro e do operador, a importância da experiência é claramente supervalorizada, na medida em que dá ao profissional credenciais cuja precariedade fica oculta sob a capa da autovalidação. Apesar de poder causar problemas que nem sempre são facilmente detectáveis, como no caso das escolhas em um projeto de RTO, esta supervalorização da experiência é duplamente amparada. Por um lado, pelas empresas que vendem sistemas de RTO, pois delega um extenso rol de decisões ao cliente, simplificando o seu trabalho de fornecedor de serviço e transferindo responsabilidade pelo desempenho. Por outro lado, ela é muitas vezes validada, implícita ou explicitamente, pela cultura corporativa da empresa usuária. Neste caso, motivações mais sutis, de natureza social e psicológica, se entrelaçam por trás de políticas de valorização de pessoal, cujas finalidades, embora nobres, podem conduzir a resultados ambíguos sob o ponto de vista da produtividade.

Ao escolher o conjunto de variáveis atualizáveis, o usuário está se colocando defronte de um problema de magnitude considerável para as limitações humanas de tomada de decisão: o de discernir, em um espaço cuja dimensão corresponda a todas as configurações e combinações de estimativas/reconciliações possíveis, aquelas escolhas que produzem desempenho esperado (em termos de sub-optimalidade e variabilidade) dentro de determinado nível de confiança estatístico, em dado conjunto de cenários de operações possíveis. Não só a dimensão deste problema é muito grande, mesmo em problemas simples, como a formato das regiões associadas a dado nível de confiança são imprevisíveis *a priori* para um problema não-linear genérico.

O problema de escolha da estrutura do RTO evidencia o seu caráter intratável quando reduzido à abordagem da “experiência” do usuário, exceção feita a problemas muito simples. Infelizmente, as conseqüências da manutenção desta abordagem não são aparentes nem metodicamente investigadas no ambiente industrial, caracterizando um problema invisível à operação cotidiana.

A escolha de variáveis reconciliáveis com “significado físico” para atuarem no duplo papel de adaptação do modelo e de ferramenta de diagnóstico deve ser evitada em

prol de maior rigor decisório. As variáveis devem ser escolhidas por meio de um crivo vinculado à função objetivo da camada principal de otimização e esta escolha deve ser estritamente pragmática, no sentido de auxiliar o desempenho global do RTO. Apesar de exemplarmente destacado por Forbes, Marlin e MacGregor [40] o fato de o melhor modelo ser aquele que sugere o conjunto de *set-points* que mais se aproxime daquele que conduz a planta à operação ótima, e não necessariamente o que melhor estima os parâmetros de relevância para o diagnóstico da planta, não é plenamente entendido. Além disso, a natureza em dupla camada de otimização costuma ser fonte de equívocos quanto à interpretação e à relevância do processo de adaptação. É comum, na prática de engenharia, qualificar, mesmo que indiretamente, o desempenho do RTO em termos da acurácia do modelo em predizer as respostas do processo em função dos estímulos, e muito esforço e análise de processo é gasto por esta causa.

O método de verificação de adequabilidade de modelos de otimização proposto por Forbes, Marlin e MacGregor [40] foi formulado para verificar se um dado conjunto de parâmetros atualizáveis permite que o processo de otimização encontre os valores das variáveis de decisão **u*** que conduzem a função objetivo ao ponto ótimo. Para tanto, além de garantir o atendimento das restrições de igualdade e de desigualdade do modelo, devem existir combinações dos valores dos parâmetros atualizáveis que garantam:

$$\begin{aligned} &\nabla_r L\big|_{x^*} = \mathbf{0} \\ &\lambda_i \geq 0\ ,\ \ \lambda_i - \textit{autovalores de } \nabla_r^2 L \end{aligned} \tag{7}$$

onde o subscrito r indica que as operações são executadas no espaço reduzido [40,41] das variáveis manipuladas da otimização (*set-points* gerados pelo RTO).

Note-se, contudo, que embora matematicamente inquestionável, há algumas limitações nesta abordagem. Esta avaliação é de validade local ao redor do ponto ótimo **u***, e não há garantia que pontos ótimos referentes a diferentes condições operacionais possam ser encontrados. O objetivo é verificar se os parâmetros escolhidos podem assumir valores que propiciem condições ao otimizador da função econômica para que se encontre **u***. Nada se fala sobre qual a influência do procedimento de atualização dos parâmetros. Por último, há as dificuldades no cálculo do gradiente e da hessiana reduzidos.

Uma alternativa de caráter eminentemente prático, embora simplista, é a de Krishnan, Barton e Perkins [38,39], que consiste em realizar a análise de sensibilidade da função lucro em relação a variações de até ± 5% dos parâmetros. A fragilidade desta abordagem é que, além de possuir caráter local, o caráter monovariável das análises impede a verificação de efeitos combinados de diferentes agrupamentos de parâmetros. Além disto, não é previsto um critério objetivo de ponto de corte dos efeitos sobre o lucro que determine a inclusão ou não de determinado parâmetro no rol daqueles atualizáveis.

Em virtude do funcionamento integrado de todos os componentes de um RTO, a análise de mudanças e decisões de projeto tomadas em cada uma de suas instâncias tem impacto no desempenho global do sistema. Por conta disto, qualquer análise sobre alternativas na estrutura de um RTO deve levar em conta o comportamento do sistema completo na presença das modificações. Métodos que observem isoladamente o efeito de cada mudança [38,39] não são capazes de prever a interação das intervenções com as demais instâncias do sistema.

Para poder discriminar estruturas alternativas para um RTO é necessário que o método de análise considere o sistema operando em malha fechada, de modo a observar os efeitos ao longo de todos os elementos do sistema. Além disto, deve ser definida uma métrica de desempenho vinculada às políticas operacionais da planta.

Há na literatura apenas duas abordagens que atendem a estes requisitos, propondo-se a lidar, de forma sistemática, com o problema de discriminar estruturas alternativas para RTO: a abordagem de Loeblein e Perkins [37,42-43], de 1996/98, sobre o trabalho original de de Hennin, Perkins e Barton [44], de 1994, por eles denominada de desvio médio do ótimo (DMO); e a abordagem de Forbes e Marlin [45], chamada de custo de projeto (CP), apresentada em 1996 e extendida por Zhang e Marlin em 2000 [4]. Muitas formulações apresentadas em [45] são idênticas às apresentadas por Pinto [46], que lidava com os custos de incertezas em problemas de otimização, ainda que este não abordasse especificamente um problema de RTO.

Apesar da similaridade do problema sobre o qual se debruçam, da parca presença deste assunto na literatura técnica da área e da contemporaneidade dos trabalhos, os autores das propostas do desvio médio do ótimo e do custo de projeto não dialogam entre si, causando a impressão de que seguem caminhos pretensamente independentes e fazendo com que as semelhanças se ocultem sob as diferentes formas de apresentar as questões. Em um esforço para remover estas diferentes apresentações e privilegiar a

análise comparativa, pode-se apresentar da seguinte forma os dois índices de desempenho propostos:

$$DMO = E\left[L(\mathbf{u}^*)\right] - E\left[L(\hat{\mathbf{u}})\right] \tag{8}$$

$$CP = L(\mathbf{u}^*) - E\left[L(\hat{\mathbf{u}})\right] \tag{9}$$

Ambas as propostas procuram criar uma medida de desempenho econômico baseada no afastamento, em relação a uma dada referência, do lucro obtido com as estimativas do ponto ótimo de operação, $\hat{\mathbf{u}}$. Diferentes estruturas conduzem a diferentes estimativas e, consequentemente, a diferentes resultados econômicos, o que permite a sua discriminação. Uma sutil diferença distingue a definição dos pontos de referência. Na Equação (9) é idealizada a existência e o conhecimento perfeito da melhor condição operacional, $\mathbf{u}^*$, e do conseqüente máximo lucro possível. Em (8) a condição de referência é considerada em um grau inferior de idealização. Ainda que os modelos do processo e as medições sejam perfeitas, o conhecimento sobre os parâmetros do modelo não o é, causando variabilidade na informação obtida a respeito do lucro máximo. Note-se, contudo, que esta incerteza a respeito dos parâmetros é constante e dada *a priori*, sendo inalterada por alterações estruturais do RTO, o que torna inócuo o trabalho de calcular $E\left[L(\mathbf{u}^*)\right]$ para o objetivo de discriminar o desempenho de diferentes sistemas. Dadas as similaridades e a forma de apresentação, ambas as formulações (Equações (8) e (9)) poderiam ser chamadas de desvio médio do ótimo.

Os dois métodos têm o objetivo de formular uma expressão analítica para os critérios de desempenho. Para tanto, propõem igualmente representar a função objetivo econômica por meio de uma expansão em série de Taylor de segunda ordem, de modo a livrar-se de particularidades e características analiticamente intratáveis que modelos não lineares possam apresentar. É a partir deste ponto que as propostas diferem.

O método CP propõe expressar o critério de desempenho em função da condição operacional estimada, $\hat{\mathbf{u}}$, deixando implícitas as fontes de variabilidade e as especificades do modelo. Desse modo, define o limite superior esperado para CP [45]:

$$CP \leq \left\| \nabla_r^2 L\big|_{\mathbf{u}^*} \right\|_2 \left\| (\mathbf{u}^* - \hat{\mathbf{u}}) \right\|_2^2 + nu \left\| \mathbf{V}_{\hat{\mathbf{u}}} \right\|_2 \tag{10}$$

onde nu=dim(**u**) é a dimensão do vetor **u**. $\mathbf{V}_{\hat{\mathbf{u}}}$, matriz covariância das predições $\hat{\mathbf{u}}$, é calculado [45] via aproximação linear da propagação das incertezas de medida pelos elementos do ciclo do RTO (medição, estimação de parâmetros, otimização) sob a suposição de erros gaussianos aditivos não correlacionados no tempo. Um desenvolvimento posterior [4] incorporou o impacto no lucro dos sucessivos ciclos do RTO até a estabilização dos valores propostos.

O DMO tem por objetivo apresentar uma expressão analítica para o critério de desempenho diretamente em função das características primárias de variabilidade, como os erros de medição e a variabilidade estocástica e determinística dos parâmetros, assim como em função do vetor de afastamento que adicione uma margem de segurança às restrições. Apesar de gerar como solução uma equação linear pronta para o uso, sem necessidade de procedimentos adicionais de otimização (embora muito extensa e, por este motivo, omitida neste texto), o DMO representa um caso menos genérico que o CP, pois já incorpora quais são as fontes de variabilidade (aditivas, gaussianas), as funções de restrição do modelo (todas linearizadas), o procedimento de estimação (mínimos quadrados, matriz de covariância diagonal, sem reconciliação de dados), o procedimento de robustez da otimização. Estas definições prévias diminuem o rol de variações estruturais do RTO que pode ser discriminado pelo método.

A Tabela 1 resume as características das métricas CP e DMO. Em nome de uma formulação analítica da medição de desempenho, ambos os métodos fazem diversas aproximações. Além disto, alguns cálculos (gradientes envolvidos no cálculo de $\mathbf{V}_{\hat{\mathbf{u}}}$) podem ser pouco práticos de serem executados em problemas maiores. Este tipo de formulação está muito vinculado ao fato de os casos estudados na literatura serem pouco complexos, sendo formulados explicitamente em termos das equações do modelo e das restrições. Normalmente são usados modelos cujos número total de equações está algumas ordens de grandeza abaixo daqueles encontrados em uso industrial. Em casos reais, na maior parte das vezes a formulação explícita das equações não está disponível pois o modelo é construído em *softwares* comerciais de simulação de processos. Tal fato é um sério obstáculo a propostas que prevejam manipulações algébricas das equações do modelo, como requerido para o cálculo de CP e DMO.

Tabela 1 - Comparação das métricas de desempenho de RTO

Características	CP	DMO
Aproximações	- função objetivo econômica: aproximada por série de Taylor de 2a ordem; curvatura constante ao longo de qualquer direção; - $\mathbf{V}_{\ddot{u}}$ calculado via aproximação linear; - restrições ativas não mudam devido à variabilidade	- função obj econômica aproximada por série de Taylor de 2a ordem; - linearização do modelo e das restrições do processo - restrições ativas não mudam devido à variabilidade
Erros das medidas	erros aditivos gaussianos não correlacionados	erros aditivos gaussianos não correlacionados prevê possibilidade de variação temporal dos parâmetros
Elementos já incluídos na estrutura do RTO	-	- Otimização em duas etapas com função objetivo pré-definida; - Uso de vetor de afastamento (Eq. 2) para as funções de restrição do modelo

Para a solução de problemas reais é mais conveniente deixar de lado as aproximações usadas nos métodos da Tabela 1 e generalizar a apresentação do problema de otimização, como será visto ao longo deste trabalho e, mais especificamente, no Capítulo 5.

1.5. Objetivos da Tese

O uso de sistemas de RTO tem se tornado cada vez mais comum na indústria. Contudo, a distância entre as discussões apresentadas na literatura e a vivência real do uso destes sistemas também tem aumentado na mesma proporção. A implementação de projetos de RTO exige vários meses de trabalho especializado e costuma produzir um sistema matemático complexo que abrange até centenas de milhares de equações. Sob a pressão do dia a dia, é muito comum que os usuários se envolvam em questões relativas à operação do sistema, ao invés de se dedicarem ao seu diagnóstico e à reflexão sobre os resultados obtidos pelas ferramentas disponíveis. Embora seja possível encontrar algumas críticas valiosas sobre implementações RTO na literatura aberta [3,47], esta discussão é geralmente apresentada em termos superficiais, o que torna difícil para os

profissionais distinguir questões relacionadas ao processo e à implementação de limitações metodológicas da abordagem RTO em si.

Na verdade, os *softwares* comerciais, considerando os grandes provedores de tecnologia, são geralmente baseados em uma abordagem muito padronizada do RTO em duas etapas, não levando em conta as melhorias colaterais da literatura técnica da área, como o *design* de excitação de entrada [48,49] ou diagnóstico automatizado [27].

Tendo estas questões em vista, o presente trabalho tem como objetivo trazer à tona questões fundamentais inerentes ao RTO em duas etapas e suas implicações no cotidiano de operação da planta. No artigo *Common Vulnerabilities of RTO Implementations in Real Chemical Processes* [50], publicado ao longo da evolução deste trabalho de doutorado, várias das questões relativas a este distanciamento são apresentadas. No presente texto, estas discussões são formalizadas, aprofundadas e desenvolvidas.

1.5.1. Contribuições Originais e Organização do Trabalho

As contribuições originais deste trabalho podem ser assim enunciadas:

- Formulação de uma descrição formal para o problema de RTO visto de forma completa, realçando seu papel no fluxo de informações produzido pelo processo;

- Discussão sobre a estabilidade de sistemas de RTO;

- Análise da utilidade dos testes de estacionariedade existentes;

- Análise crítica de dados reais de dados industriais de RTO;

- Proposição de estruturas de RTO segundo um critério exaustivo amparado por métricas pragmáticas.

Estes temas são organizados da seguinte forma:

O Capítulo 2 formula o problema de RTO, estabelece as noções de incompletude, corrupção e processamento incorreto da informação como a base das discussões sobre a degradação de desempenho do RTO. Também apresenta o problema da configuração da estrutura de um sistema de RTO, definindo as regras e restrições de todas as possíveis estruturas que um sistema pode apresentar.

O Capítulo 3 detalha as instâncias de adaptação do modelo e otimização financeira à luz da formalização proposta e exemplifica as causas da variabilidade introduzida no processo pela presença do RTO, realçando a possível instabilidade derivada de suas ações. Além disto, revê os principais métodos de detecção de estacionariedade, mostrando suas fragilidades de desempenho no contexto de um sistema de RTO, apontando para soluções alternativas voltadas para o conceito de adequabilidade e utilidade, em oposição ao conceito de estacionariedade comumente usado.

O Capítulo 4 apresenta a arquitetura típica de sistemas de RTO comerciais, mostrando as soluções tipicamente usadas para detecção de estacionariedade e para adaptação do modelo. Discute a conveniência destas soluções sob os conceitos apresentados nos Capítulos 2 e 3 e apresenta a análise de dados reais produzidos pelo RTO de uma unidade de destilação de petróleo ao longo de 1000 ciclos de otimização em tempo real em malha fechada.

O Capítulo 5 traz um estudo de caso de definição de estrutura de um sistema de RTO para o processo reacional de Williams-Otto. As regras enunciadas no Capítulo 2 são aplicadas para a enumeração de todas as estruturas possíveis para três versões do problema. Em conjunto, são estabelecidas métricas para discriminação da conveniência associada à escolha de cada estrutura. Em seguida, o RTO configurado tem seu desempenho quantificado e comparado relativamente às diversas versões do problema.

2. Formulação do Problema do RTO

Neste capítulo é feito o desenvolvimento formal do arcabouço matemático do problema de otimização em tempo real sob o ponto de vista do fluxo de informações associado à sua inserção em um processo. São descritas as diversas instâncias que o compõe e as vulnerabilidades inerentes ao RTO em sua tarefa de continuamente apropriar-se de informações da instrumentação e formular decisões ótimas. São também elencadas as escolhas associadas à configuração de sua estrutura e formuladas as regras que definem e limitam estas escolhas. Para tal, é desenvolvida uma simbologia que atenda estes objetivos, partindo da representação dinâmica do processo, **ZZ**, até o caso particular **Z**, que reúne as informações descritoras dos estado estacionário.

2.1. Definição de Processo

Define-se o processo $\mathcal{P}$ (Equação 11) como a região no espaço $R^{\dim(\mathbf{ZZ})}$ descrita pelas relações funcionais $\boldsymbol{f}$ e $\boldsymbol{g}$ (Equações 12 e 13), onde **ZZ** consiste no conjunto finito de todas as informações quantitativas que descrevem, de forma completa, o recorte de interesse da realidade física em estudo. Na Equação 12, cada vetor $\mathbf{fx}_i$ contém o conjunto de índices que assinalam os elementos de **ZZ** que fazem parte da i-ésima equação em $\boldsymbol{f}$. Raciocínio análogo descreve cada vetor que compõe **gx** na Equação 12. Como exemplo, se $\mathbf{fx}_2 = [5\ \ 7\ \ 14]$, sabe-se que a segunda equação do sistema $\boldsymbol{f}$ faz uso da quinta, sétima e décimo-quarta entidades (variáveis) contidas em **ZZ**. No presente texto, a função dim(●), aplicada a um vetor, matriz ou conjunto, representa o número de elementos de cada uma das dimensões do operando.

$$\mathcal{P} = \{\mathbf{ZZ} \in R^{\dim(\mathbf{ZZ})} \mid \boldsymbol{f} \wedge \boldsymbol{g}\} \tag{11}$$

$$\boldsymbol{f}: f_i(\mathbf{ZZ}(\mathbf{fx}_i)) = 0,\ i = 1..\dim(\boldsymbol{f}),\ \mathbf{fx}_i \subset \{1,2,...,\dim(\mathbf{ZZ})\} \tag{12}$$

$$\boldsymbol{g}: g_i(\mathbf{ZZ}(\mathbf{gx}_i)) \leq 0,\ i = 1..\dim(\boldsymbol{g}),\ \mathbf{gx}_i \subset \{1,2,...,\dim(\mathbf{ZZ})\} \tag{13}$$

Em virtude da existência do conjunto de relações funcionais ***f***, o conjunto **ZZ** é completamente descrito a partir de um conjunto menor de informações necessárias, **ZZ**(**in**), de dimensão dim(**ZZ**)-dim(***f***), que define, a partir da transformação *Tproc*, o conjunto de informações conseqüentes, **ZZ**(**out**), de acordo com (14):

$$\begin{aligned} &Tproc: \mathsf{R}^{\dim(\mathbf{ZZ})-\dim(\boldsymbol{f})} \rightarrow \mathsf{R}^{\dim(\boldsymbol{f})} \\ &\mathbf{ZZ}(\mathbf{in}) \mapsto \mathbf{ZZ}(\mathbf{out}) \end{aligned} \tag{14}$$

Desta forma, **ZZ** pode ser equivalentemente representado em diversos espaços de dimensão reduzida, dim(**in**), provenientes de um elenco de subconjuntos de representações possíveis, **In**, (Equação15):

$$\begin{aligned} &\mathbf{In}\{i\} = \mathbf{ZZ}(\mathbf{in}_i) \\ &\mathbf{in}_i \subset \{1,2,...,\dim(\mathbf{ZZ})\}, \\ &\dim(\mathbf{in}_i) = \dim(\mathbf{ZZ}) - \dim(\boldsymbol{f}) \end{aligned} \tag{15}$$

Sob um ponto de vista genérico, quaisquer das bases de representação contidas em **In** que não representem combinações lineares e que não representem incoerências que impeçam a solução são suficientes e equivalentes para definir **ZZ**. Contudo, nem todos os subconjuntos de **ZZ** de dimensão dim(**ZZ**) – dim(***f***) fazem parte de **In**. A restrição genérica de colinearidade apresentada na Equação (16) impõe restrições para o total de representações mínimas contidas em **In**, de acordo com (17).

$$\begin{aligned} &f_i(\mathbf{ZZ}(\mathbf{fx}_i)) = 0, \ i = 1..\dim(\boldsymbol{f}) \\ &\mathbf{fx}_i \setminus \mathbf{in} \neq \phi, \ \forall i \end{aligned} \tag{16}$$

$$\dim(\mathbf{In}) < \frac{\dim(\mathbf{ZZ})!}{(\dim(\mathbf{ZZ}) - \dim(\boldsymbol{f}))!\dim(\boldsymbol{f})!} \tag{17}$$

Considerando-se o problema genérico, é conveniente particionar as relações funcionais em ***f***(**ZZ**) nos conjuntos de equações algébricas e diferenciais (18, 19,20), que se distinguem pela dependência explícita com a derivada das variáveis de estado. A Equação (21) indica de que modo os elementos de **ZZ** assinalados pelos índices **ds** se

relacionam com as variáveis de estado, assinalados pelos índices **s**, de acordo com a moldura genérica representada pelas funções **q.**

$$f = \left\{ f_{a\lg},\ f_{dif} \right\} \tag{18}$$

$$f_{a\lg,\,i} = \left\{ f_i(\mathbf{ZZ}(\mathbf{fx}_i)) = 0 \mid \mathbf{ds} \cap \mathbf{fx}_i = \phi \right\} \tag{19}$$

$$f_{dif,\,i} = \left\{ f_i(\mathbf{ZZ}(\mathbf{fx}_i)) = 0 \mid \mathbf{ds} \cap \mathbf{fx}_i \neq \phi \right\} \tag{20}$$

$$\left\{ \mathbf{s}, \mathbf{ds} \subset \{1,\ldots \dim(\mathbf{ZZ})\} \mid \ \mathbf{ZZ}(\mathbf{ds}(i)) = \frac{d\, q_i(\mathbf{ZZ}(\mathbf{j}))}{dt},\ i \in \mathbf{ds},\ \mathbf{j} \subset \mathbf{s} \right\} \tag{21}$$

Prosseguindo com as consequências da partição entre equações algébricas e diferenciais, é conveniente discriminar os elementos de **ZZ** que não dependem dos seus valores pregressos (**II**), daqueles que dependem (**OO**), de acordo com as Equações (22-28). Note-se que, embora as variáveis em **OO** possam, sob o ponto de vista matemático, fazer parte das variáveis necessárias **in** em (14), em termos físicos elas representam variáveis que não podem ser diretamente manipuladas para a condução do processo. Assim sendo, excluindo-se a consideração de processos não causais e admitindo-se puramente o ponto de vista da realidade física subjacente, **II** contém os elementos de entrada (Equação 23) e **OO** está contido nos elementos de saída do processo (Equação 24). A partição de **II** entre as variáveis **I** e **τ**, de acordo com o mostrado na Equação (25), tem por finalidade evidenciar **τ**, que representa um conjunto de informações ao qual ***f*** é insensível caso **dO** seja nulo (Equação 29).

$$\{\mathbf{ZZ}\} = \left\{ \mathbf{II},\ \mathbf{OO} \right\} \tag{22}$$

$$\mathbf{II} \supseteq \{\mathbf{ZZ}(\mathbf{in})\} \tag{23}$$

$$\mathbf{OO} \subseteq \{\mathbf{ZZ}(\mathbf{out})\} \tag{24}$$

$$\mathbf{II} = \left[\mathbf{I}^{\mathrm{T}}, \boldsymbol{\tau}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{25}$$

$$\mathbf{OO} = \left[\mathbf{O}^{\mathrm{T}}, \mathbf{dO}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{26}$$

onde

$$\mathbf{O} = \mathbf{ZZ}(\mathbf{s}) \tag{27}$$

$$\mathbf{dO} = \mathbf{ZZ}(\mathbf{ds}) \tag{28}$$

$$\mathbf{OO} \in \mathrm{R}^{2\dim(\mathbf{s})}$$

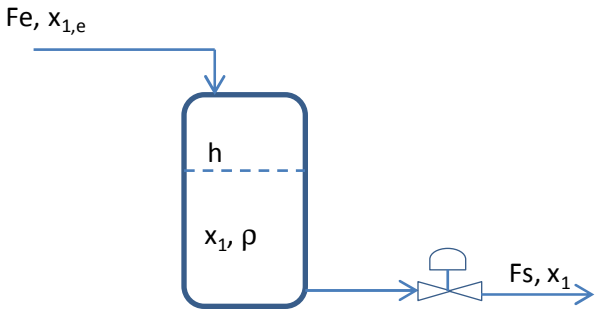
$$\mathbf{dO} = \mathbf{0} \;\Rightarrow\; \frac{d\boldsymbol{f}_i}{d\boldsymbol{\tau}(j)} = 0, \;\; \forall i, j \tag{29}$$

De acordo com a presente representação, o caso particular de um processo que evolua (ou que seja observado) apenas pela sucessão de diversos estados estacionários é completamente descrito pelo subconjunto **Z** (Equação 30). Embora a lembrança da realidade física pertinente remeta à distinção dos conjuntos **I** e **O** como sendo respectivamente referidos às entradas e saídas do processo, neste caso tais distinções são matematicamente inócuas pois ambos os conjuntos relacionam-se apenas por meio das relações algébricas, $\boldsymbol{f}_{\mathrm{alg}}$, que coincidem com o conjunto $\boldsymbol{f}$.

$$\mathbf{Z} = \left\{ \mathbf{x} \,|\, \mathbf{x} \subseteq \{\mathbf{I}, \mathbf{O}\} \right\} \tag{30}$$

O exemplo 1, mostrado a seguir, pode ser útil para esclarecer a simbologia adotada. Considere o reator cujo processo é descrito pelas variáveis **ZZ** (Equação 31), restritas pelas relações funcionais $\boldsymbol{f}$ (Equação 32) e $\boldsymbol{g}$ (Equação 33). As variáveis **O**, **dO**, e seus respectivos índices em **ZZ**, **s** e **ds**, são formulados de acordo com as Equações (21,27,28) e representadas nas Equações (34-36). As variáveis **II** são, por consequência, definidas pela Equação (37).

Exemplo 1 – Descrição de um processo químico e partição das variáveis de estado



$$\{\mathbf{ZZ}\} = \left\{ Fe, x_{1,e}, c, k, A, \rho, Fs, h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\} \tag{31}$$

$$\boldsymbol{f}: \begin{cases} \boldsymbol{f}_{alg}: \{ Fs = c\sqrt{h} \\ \boldsymbol{f}_{dif}: \begin{cases} A\rho \dfrac{dh}{dt} = Fe - Fs \\ A\rho \dfrac{dhx_1}{dt} = Fe\,x_{1,e} - Fs\,x_1 - kx_1 Ah \end{cases} \end{cases} \tag{32}$$

$$\boldsymbol{g}: \begin{cases} h - h_{\max} \le 0 \\ -h < 0 \end{cases} \tag{33}$$

$$\{\mathbf{O}\} = \{h, x_1\}, \ \mathbf{s} = [8, 9] \tag{34}$$

$$\{\mathbf{dO}\} = \left\{ \frac{dh}{dt}, \frac{d(hx_1)}{dt} \right\}, \ \mathbf{ds} = [10, 11]$$

(35)

$$q_1(x_1, h) = h \ ; \ q_2(x_1, h) = x_1.h \tag{36}$$

$$\{\mathbf{II}\} = \{\mathbf{ZZ}\} \setminus \{\mathbf{OO}\} = \left\{ \underbrace{Fe, x_{1,e}, c, k, Fs}_{\mathbf{I}}, \underbrace{A, \rho}_{\tau} \right\} \tag{37}$$

2.2. Obstáculos ao conhecimento verdadeiro do Processo

2.2.1. Incompletude das Informações

Em dado horizonte de estados consecutivos {$\mathbf{ZZ}_j$, $\mathbf{ZZ}_{j+1}$...}assumidos pelo processo, aquelas variáveis que não agregam informação nova ao conhecimento dos estados são assinaladas pelos índices **fix**, conforme a Equação (38), em oposição àquelas assinaladas pelos índices **var (**Equação (39)), cuja premissa de variabilidade implica na modificação do conteúdo de informação em **ZZ** .

$$\mathbf{fix} = \left\{ \mathbf{x} |\ \mathbf{ZZ}_j(\mathbf{x}) = \mathbf{ZZ}_0(\mathbf{x}),\ \forall\, j \right\} \tag{38}$$

$$\mathbf{ZZ}(\mathbf{var}) = \mathbf{ZZ} \setminus \mathbf{ZZ}(\mathbf{fix}) \tag{39}$$

Em virtude da existência da estrutura $\boldsymbol{f}$, limitações adicionais são impostas à construção do conjunto **In**. Por exemplo, os elementos **var** nunca formarão um subconjunto de **in** (Equação 40), na medida em que a Equação (16) for verdadeira.

Tomando como exemplo o caso mais simples, em que dim(**var**)=2, a conseqüência da Equação (40) é que estas variáveis estarão particionadas de modo que $\dim(\mathbf{in} \cap \mathbf{var}) = \dim(\mathbf{out} \cap \mathbf{var}) = 1$, ou seja, se há uma variável necessária (entrada) que se modifica deve haver ao menos uma variável consequente (de saída) que também se modifique.

Os limites para a partição das variáveis **ZZ**(**var**) no conjunto de variáveis necessárias, **ZZ**(**in**), e de variáveis conseqüentes, **ZZ**(**out**), podem ser expressos, de modo genérico, pelas Equação (41) e (42), que são fundamentadas na possibilidade de existência dos casos extremos em que um único elemento de $\mathbf{in} \cap \mathbf{var}$ afeta todas as variáveis conseqüentes (Equação 43) ou na possibilidade de que todos os elementos de $\mathbf{in} \cap \mathbf{var}$ afetem apenas uma das relações em $\boldsymbol{f}$ (Equação 44).

$$\mathbf{var} \not\subset \mathbf{in} \tag{40}$$

$$1 \le \dim(\mathbf{out} \cap \mathbf{var}) \le min\left(\dim(\mathbf{var}) - 1,\ \dim(\boldsymbol{f})\right) \tag{41}$$

$$\dim(\mathbf{var}) - min\left(\dim(\mathbf{var}) - 1,\ \dim(\boldsymbol{f})\right) \leq \dim(\mathbf{in} \cap \mathbf{var}) \leq \dim(\mathbf{var}) - 1 \quad (42)$$

$$\exists! k:\ k = (\mathbf{in} \cap \mathbf{var}),\ k \in \mathbf{fx}_i,\quad \forall i \in \{1,..,\dim(\boldsymbol{f})\} \quad (43)$$

$$\exists! i:\ (\mathbf{in} \cap \mathbf{var}) \subset \mathbf{fx}_i \quad (44)$$

As Equações (15-17) formulam o conjunto de escolhas para a representação mínima de **ZZ** sob o ponto de vista de sua equivalência matemática para o mapeamento (Equação 14), estruturado nas relações funcionais ***f***. Contudo, a diferenciação dos elementos de **ZZ** em função da informação a eles associada se ampara em um conhecimento exterior à representação matemática, vinculado à realidade física do processo descrito. As implicações (Equações 40-42) derivadas desta diferenciação sobre a partição desses elementos em variáveis necessárias e conseqüentes assinalam o uso de um conhecimento *a priori*, não necessariamente sujeito a revalidação periódica ao longo da operação.

Neste ponto, torna-se necessário formular uma hipótese fundamental para o presente trabalho: a de que existe uma verdade determinística que representa o objeto em estudo, dado pelos elementos de $\mathcal{P}$, que pode ser referenciada apenas em termos idealizados, sendo inacessível sua verificação inequívoca. Em contrapartida, o que se pode dispor é uma versão alternativa, $\mathcal{P}m$, baseada no conhecimento aparente de $\mathcal{P}$ (Equação 45). Todas as transformações, diagnósticos e tomadas de decisão somente podem ser referidas a $\mathcal{P}m$, que se baseia na estrutura ***fm*** e ***gm*** (análogos conhecidos de ***f*** e ***g***) e em **ZZm** (análogo conhecido de **ZZ**).

$$\mathcal{P}m = \{\mathbf{ZZm} \in \mathbb{R}^{\dim(\mathbf{ZZ})} \mid \boldsymbol{fm} \wedge \boldsymbol{gm}\} \quad (45)$$

O conhecimento aparente, $\mathcal{P}$m, é formado a partir da composição de dois grupos de informação (Equação 46), de naturezas distintas: *Iod*, que representa as informações obtidas por observação direta; ou seja, que são adquiridas independentemente do conhecimento das relações estruturais ***fm*** e ***gm***; *Iap*, que representa as informações tomadas *a priori* a respeito do processo, e não sujeitas a posterior revalidação.

$$\{Iap, Iod\} \rightarrow \mathbf{ZZm} \tag{46}$$

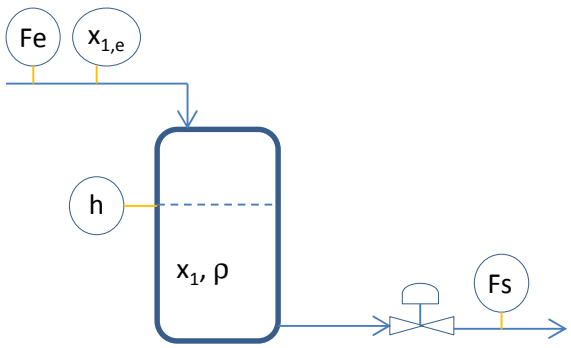
As observações diretas representam o conjunto **ZZa** (47**)**, que espelha o conteúdo da verdade, embora com duas restrições importantes: ainda que **ZZ** evolua de forma contínua entre seus sucessivos estados ao longo do tempo, as observações **ZZa** incorporam a incapacidade de distinguir entre dois estados separados por um intervalo menor que T_{am} (Equação 48); além disto, **ZZa** não incorpora todas as informações contidas em **ZZ**, mas apenas o subconjunto **ms**. A relação de **ZZa** com **ZZ** é representada, de forma ideal, na Equação 49, onde $\overline{\mathbf{ms}}$ representa os elementos de **ZZa** complementares a **ms**. Na Seção 2.2.3 esta relação será descrita de forma mais abrangente. O exemplo 2 mostra estes conceitos aplicados ao caso estudado no exemplo 1 da página 26.

$$Iod = \{\mathbf{ZZa}\} \tag{47}$$

$$\mathbf{ZZa}_j = \mathbf{ZZ}_{jT_{am}}, \quad j \in \mathbb{N} \tag{48}$$

$$\begin{aligned} &\mathbf{ZZa(ms)} = \mathbf{ZZ(ms)}; \\ &\mathbf{ZZa}(\overline{\mathbf{ms}}) = \mathbf{0} \\ &\mathbf{ms} \subseteq \{1, 2, \ldots \dim(\mathbf{ZZ})\} \end{aligned} \tag{49}$$

Exemplo 2: Informações obtidas por observação direta



ZZm=ZZ, ***fm*** = ***f***

$$\{\mathbf{ZZm}\} = \left\{ Fe, x_{1,e}, c, k, A, \rho, Fs, h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\}$$

$$\{\mathbf{ms}\} = \{ Fe, x_{1,e}, Fs, h \}, \quad \mathbf{ms} = [1, 2, 7, 8]$$

$$\mathbf{ZZa}_j = [\mathbf{ZZ}_{jTam}(1), \mathbf{ZZ}_{jTam}(2), \ 0, 0, 0, 0, \ \mathbf{ZZ}_{jTam}(7), \mathbf{ZZ}_{jTam}(8), 0, 0, 0]^T$$

As informações apriorísticas contemplam o conhecimento do processo em sua condição de referência, $\mathbf{ZZm}_0$ (também referida como condição nominal, condição de projeto, caso-base, dentre outras denominações), assim como a natureza suposta para as relações funcionais, ***fm*** e ***gm*** (Equação 50). Consistem em um conjunto heterogêneo quanto à origem, reunindo informações diretas preliminares, hipóteses verificáveis experimentalmente ou não, e regras e expectativas empíricas agregadas com diferentes graus de subjetividade.

$$Iap = \{\mathbf{ZZm}_0, \boldsymbol{fm}, \boldsymbol{gm}\} \tag{50}$$

Cada elemento do conjunto de escolhas **In**, referido em (15,17) induz a diferentes recortes do conjunto de informações disponível. Através do mapeamento (51), cada recorte produzirá, de forma correspondente, o conjunto-base de representação do processo. Este mapeamento se apóia no conjunto expandido de relações funcionais $\boldsymbol{f}_{\text{sis}}$ (52), que inclui, além das relações ***fm***, as relações de medição $\boldsymbol{f}_{med}$, que incorporam as informações obtidas por observação direta, *Iod*; as relações de atribuição, $\boldsymbol{f}_{atr}$, que incorporam as informações inseridas a priori, *Iap;* e as condições iniciais, $\boldsymbol{f}_{ini}$, que dão conta dos graus de liberdade dinâmicos.

$$\begin{aligned} T\mathcal{P}m &: \mathbb{R}^{\dim(\mathbf{ZZ})-\dim(\boldsymbol{f})} \rightarrow \mathbb{R}^{\dim(\mathbf{ZZ})-\dim(\boldsymbol{f})} \\ \{Iod, Iap\} &\mapsto \mathbf{ZZm}(\mathbf{in}) \end{aligned} \tag{51}$$

$$\boldsymbol{f}_{sis} = \boldsymbol{fm} \cup \boldsymbol{f}_{med} \cup \boldsymbol{f}_{atr} \cup \boldsymbol{f}_{ini} \tag{52}$$

As relações de medição $\boldsymbol{f}_{med}$ concatenam em **ZZm** o subconjunto $\mathbf{ms}^{-}$ dos elementos obtidos por observação direta que estão incluídos no elenco de variáveis necessárias, de acordo com (53). O conjunto dos elementos medidos mas não incluídos

nas atribuições em ***fmed***, **ms\ms⁻**, tem seu uso vinculado ao contexto da adaptação do modelo e está descrito na Seção 2.3.

$$\begin{aligned} &\boldsymbol{f}_{med}:\ \mathbf{ZZm}_j(\mathbf{ms}^-) = \mathbf{ZZa}_j(\mathbf{ms}^-) \\ &\mathbf{ms}^- = \mathbf{ms} \cap \mathbf{in} \\ &0 \le \dim(\mathbf{ms}^-) \le \min\left(\dim(\mathbf{ms}), \dim(\mathbf{in})\right) \\ &\dim(\boldsymbol{f}_{med}) = \dim(\mathbf{ms}^-) \end{aligned} \tag{53}$$

As relações de atribuição, $\boldsymbol{f}_{atr}$, são responsáveis por incorporar o conjunto de informações a priori em **ZZm**, por intermédio dos elementos referenciados por **atr** (54). Para o caso de um sistema dinâmico (**ds=ϕ**), atribuições complementares, referentes às condições iniciais são assinaladas por meio das relações $\boldsymbol{f}_{ini}$, em (55).

O vetor **atr** possui dimensão compatível com a existência de grau de liberdade (GL) nulo, condição requerida para o mapeamento *Tproc* (14) ser efetuado (56, 57, 58).

$$\begin{aligned} &\boldsymbol{f}_{atr}:\mathbf{ZZm}_j(\mathbf{atr}) = \mathbf{ZZm}_0(\mathbf{atr}) \\ &\mathbf{atr} = \mathbf{in} \setminus \mathbf{ms}^- \\ &\dim(\boldsymbol{f}_{atr}) = \dim(\mathbf{atr}) \end{aligned} \tag{54}$$

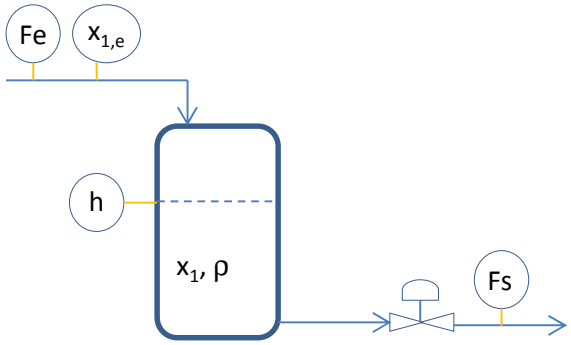
$$\begin{aligned} &\boldsymbol{f}_{ini}:\mathbf{ZZm}_0(\mathbf{s}) = \mathbf{v_0} \\ &\dim(\boldsymbol{f}_{\mathbf{ini}}) = \dim(\mathbf{s}) = \dim(\mathbf{O}) \end{aligned} \tag{55}$$

$$GL = \dim(\mathbf{ZZm}) - \left(\underbrace{\dim(\boldsymbol{f}) + \dim(\boldsymbol{f}_{atr}) + \dim(\boldsymbol{f}_{med}) + \dim(\boldsymbol{f}_{ini})}_{\dim(\boldsymbol{f}sis)} \right) \tag{56}$$

$$Tproc(Tpm(\mathbf{ZZm}_0, \mathbf{ZZa})) = \mathbf{ZZm}(\mathbf{out}) \tag{57}$$

$$\mathbf{ZZ}_j \xrightarrow{Tmed} \left\{ \begin{matrix} \mathbf{ZZm}_0(\mathbf{atr}) \\ \mathbf{ZZa}_j(\mathbf{ms}) \end{matrix} \right\} \xrightarrow{Tpm} \mathbf{ZZm}_j(\mathbf{in}) \xrightarrow{Tproc} \mathbf{ZZm}_j(\mathbf{out}) \tag{58}$$

Exemplo 3: Funções de medição, atribuição e de inicialização



$$\{\mathbf{ZZm}\} = \left\{ Fe, x_{1,e}, c, k, A, \rho, Fs, h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\}$$

$$\{\mathbf{ZZ(ms)}\} = \{ Fe, x_{1,e}, Fs, h \}$$

Funções de Medição – inserção das informações obtidas por informação direta

se $\{\mathbf{in}\} = \{Fe, x_{1,e}, c, k, A, \rho\}$

$$\mathbf{ms}^{-} = \mathbf{ms} \cap \mathbf{in} \Rightarrow \{\mathbf{ZZ}(\mathbf{ms}^{-})\} = \{Fe, x_{1,e}\}$$

$\mathbf{ms}^{-} = [1, 2]^{T}$

$$f_{med} \begin{cases} \mathbf{ZZm}_j(1) = \mathbf{ZZa}_j(1) \\ \mathbf{ZZm}_j(2) = \mathbf{ZZa}_j(2) \end{cases}$$

Funções de atribuição – inserção das informações apriorísticas

$\mathbf{atr} = \mathbf{in} \setminus \mathbf{ms}^{-} \Rightarrow \{\mathbf{ZZ(atr)}\} = \{c, k, A, \rho\}$

$\mathbf{atr} = [3,4,5,6]^{T}$

$$f_{atr} \begin{cases} \mathbf{ZZm}_j(3) = \mathbf{ZZm}_0(3) \\ \mathbf{ZZm}_j(4) = \mathbf{ZZm}_0(4) \\ \mathbf{ZZm}_j(5) = \mathbf{ZZm}_0(5) \\ \mathbf{ZZm}_j(6) = \mathbf{ZZm}_0(6) \end{cases}$$

Funções de inicialização – inserção das condições iniciais

$\mathbf{s} = [8,9]^{\mathrm{T}}, \quad \Rightarrow \{\mathbf{ZZ(s)}\} = \{h, x_1\}$

$$\boldsymbol{f}_{med}\begin{cases}\mathbf{ZZm}_0(8) = h(0)\\ \mathbf{ZZm}_0(9) = x_1(0)\end{cases}$$

$$GL = \dim(\mathbf{ZZm}) - \left(\underbrace{\dim(\boldsymbol{f}) + \dim(\boldsymbol{f}_{atr}) + \dim(\boldsymbol{f}_{med}) + \dim(\boldsymbol{f}_{ini})}_{\dim(\boldsymbol{f}sis)}\right)$$

$$GL = 11 - \left(\underbrace{3+4+2+2}_{\dim(\boldsymbol{f}sis)}\right) = 0$$

É importante registrar que existem outras possibilidades de associações de regras constitutivas do conjunto de informações *a priori* que não necessariamente a suposição de constância dos valores da condição inicial ao longo do cenário de operação. Como mostrado em (59), as relações de atribuição poderiam ser formuladas a partir de um elenco mais amplo, incluindo a possível vinculação do valor atribuído atual a uma função do histórico pregresso de variações por meio de relações funcionais baseadas em expectativas empíricas sobre a variabilidade da informação.

$$\boldsymbol{f}_{atr}:\mathbf{ZZm}_j(\mathbf{atr}) = \mathbf{a}$$

$$\mathbf{a}(i) \in \{\mathbf{ZZm}_0(\mathbf{atr}(i)),\ \mathbf{ZZm}_{j-1}(\mathbf{atr}(i)),\ fnc(\mathbf{ZZm}_{j-n,\ldots,j}(\mathbf{atr}(i)), \ldots\} \qquad (59)$$

Um problema fundamental que se apresenta é que, embora exista um conjunto de informações (**var**) que é dependente da posição temporal, a capacidade disponível para incorporar novas informações por observação direta é limitada ao subconjunto **ms**, previamente estabelecido à margem das reais configurações de **var**. De modo que seja possível que $\mathcal{P}m$ espelhe a real natureza de $\mathcal{P}$, é necessário que três condições sejam atendidas:

- a configuração de **var** deve ser tal que seja possível, dentre as possibilidades de formulação das escolhas de **in** e **ms**$^{-}$, que a relação (60) seja verdadeira;

- as funções de atribuição correspondam a elementos invariantes (61), o que é uma consequência da condição anterior;

- as relações de atribuição tenham sido corretamente assinaladas, conforme a Equação (62).

$$(\mathbf{in} \cap \mathbf{var}) \subseteq \mathbf{ms}^{-} \quad \Leftrightarrow (\mathbf{in} \cap \mathbf{var}) \setminus \mathbf{ms}^{-} = \phi \tag{60}$$

$$\mathbf{atr} \subseteq \mathbf{fix} \tag{61}$$

$$\mathbf{ZZm}_0(\mathbf{atr}) = \mathbf{ZZ}_0(\mathbf{atr}) \tag{62}$$

Idealmente, as informações consequentes, **ZZ**(**out**), deveriam ser induzidas pelo mapeamento *Tproc* (63) a partir das informações necessárias **ZZ**(**in**) que atendessem aos requisitos apresentados na Equação (64).

Contudo, em virtude da incompletude de informações, a representação do processo, **Zm**, atribui, *a priori*, valores às variáveis não medidas (**in****ms**$^{-}$), que não necessariamente correspondem àquelas invariáveis (**in****var**). Esta distinção tem um significado profundo na medida em que quaisquer conjuntos **ZZm**(**in**) serão formados por projeções de **ZZ**(**in**) no hiper-plano W das informações a priori, conforme expresso na Equação (65).

$$Tproc\ (\mathbf{ZZ}(\mathbf{in})) \rightarrow \mathbf{ZZ}(\mathbf{out}) \tag{63}$$

$$\begin{cases} \mathbf{ZZ}(\mathbf{in} \setminus \mathbf{var}) = \mathbf{ZZ}_0(\mathbf{in} \setminus \mathbf{var}) \\ \mathbf{in} \setminus \mathbf{var} \subseteq \mathbf{fix} \end{cases} \tag{64}$$

$$\begin{aligned} &\mathbf{ZZm}(\mathbf{in}) = \mathrm{Proj}_W(\mathbf{ZZ}(\mathbf{in})) \\ &W : \mathbf{ZZm}(\mathbf{in} \setminus \mathbf{ms}^{-}) = \mathbf{ZZm}_0(\mathbf{in} \setminus \mathbf{ms}^{-}) \end{aligned} \tag{65}$$

Os elementos que compõem o conjunto de informações necessárias **ZZ**(**in**) e que não estão incorporados nas funções de medição, ($\mathbf{in} \setminus \mathbf{ms}^{-}$), definem o plano de projeção em (65). Estes elementos são assinalados por intermédio das funções de atribuição (54).

Em condições ideais de acesso à informação (atendimento das Equações 60-62), **ZZ**(**in**) seria coincidente com sua projeção. Para que isto ocorresse, o conhecimento *a priori* deveria integralmente expressar informações sobre variáveis contidas no conjunto de informações *a priori* verdadeiras, **apv**, em (66). No mundo real, contudo, o tipo de informação contido nas relações de atribuição é mais variado (67), contemplando também estímulos ocultos, **ocu**, inacessíveis à observação direta (68), e hipóteses falsas, **apf** (69). Estes fatos distinguem de forma permanente o conjunto de informações consequentes, **ZZ**(**out**) (Equação 63), de seu análogo conhecido, **ZZm**(**out**) (Equação 70).

$$\mathbf{apv} = \left\{ \mathbf{x} \mid \mathbf{x} \subseteq (\mathbf{fix} \cap \mathbf{atr}),\ \mathbf{ZZm}_0(\mathbf{x}) = \mathbf{ZZ}_0(\mathbf{x}) \right\} \tag{66}$$

$$\mathbf{atr} = \mathbf{in} \setminus \mathbf{ms}^{-} = \left\{ \mathbf{apv}, \mathbf{apf}, \mathbf{ocu} \right\} \tag{67}$$

$$\mathbf{ocu} = (\mathbf{in} \cap \mathbf{var}) \setminus \mathbf{ms}^{-} \tag{68}$$

$$\mathbf{apf} = \left\{ \mathbf{x} \mid \mathbf{x} \subseteq (\mathbf{fix} \cap \mathbf{atr}),\ \mathbf{ZZm}_0(\mathbf{x}) \neq \mathbf{ZZ}_0(\mathbf{x}) \right\} \tag{69}$$

$$\begin{aligned} &\mathit{Tproc}(\mathbf{ZZm}(\mathbf{in})) \rightarrow \mathit{Tproc}(\mathbf{ZZm}(\mathbf{out})) \\ &\mathbf{ZZm}(\mathbf{in}) = \left\{ \mathbf{ZZm}(\mathbf{ms}^{-}) = \mathbf{ZZa}(\mathbf{ms}^{-}),\ \mathbf{ZZm}(\mathbf{in} \setminus \mathbf{ms}^{-}) = \mathbf{ZZm}_0(\mathbf{in} \setminus \mathbf{ms}^{-}) \right\} \end{aligned} \tag{70}$$

Dadas as condições de incompletude expressas pela existência dos conjuntos **ocu** e **apf** em (67), o problema preliminarmente colocado até este ponto consiste em prever uma forma de seleção dos elementos candidatos a variáveis necessárias, contidos no conjunto **In**, de tal modo que estes elementos sejam diferenciados em função das condições de contorno impostas pela formulação do problema, conforme a Equação (71). Estas condições consistem no conjunto de variáveis, **var**, que contribuem com informação nova sobre o estado do processo; o conjunto de informações obtido por observação direta, **ms**, e o conhecimento do estado preliminar do processo, $\mathbf{ZZm}_0$.

$$\{\mathbf{var}, \mathbf{ms}, \mathbf{ZZm}_0\} \xrightarrow{?} \mathbf{in} \qquad (71)$$

Exemplo 4 - Condições de acesso à informação

Partindo do sistema instrumentado como no exemplo 3:

$$\{\mathbf{ZZ(ms)}\}=\{Fe, x_{1,e}, Fs, h\}$$

Cuja condição de partida, $\mathbf{ZZ}_0$*, e seu análogo conhecido,* $\mathbf{ZZm}_0$*, são dados por:*

	Fe	x_{1e}	c	k	A	r	Fs	h	x_1	dh/dt	$d(hx_1)/dt$
$\mathbf{ZZ_0}$	1	1	1	1	1	1	1	1	1	0	-1
$\mathbf{ZZm_0}$	1	1	0.9	1	1	1	0.9	1	1	0.1	-1.1

Para fins de estudo, dois cenários de variabilidade serão considerados:

$$v1 = \{\mathbf{var}\}=\{Fe, x_{1,e}, k, Fs, h, x_1\}$$

$$v2 = \{\mathbf{var}\}=\{Fe, x_{1,e}, c, Fs, h, x_1\}$$

Dependendo da escolha do conjunto de variáveis necessárias, ***in*** *e das variabilidade de condições ao qual o sistema está submetido,* ***var,*** *podem ser negadas as condições ideais de acesso à informação. No primeiro cenário, v1, estas condições são violadas para ambos os conjuntos de variáveis necessárias apresentados, pois* $\boldsymbol{atr} \neq \boldsymbol{apv}$*. No cenário v2, tais condições são atendidas, e* $\boldsymbol{apf} = \boldsymbol{ocu} = \{\}$*.*

	v1		*v2*
{**in**}	{Fe,$x_{1,e}$, c, k, A, ρ}	{Fe,$x_{1,e}$, c, Fs, A, ρ}	{Fe,$x_{1,e}$, k, Fs, A, ρ}
{ $\mathbf{ms}^-$ }	{ Fe, $x_{1,e}$ }	{ Fe, $x_{1,e}$, Fs }	{ Fe, $x_{1,e}$, Fs }
{**atr**}	{ c, k, A, ρ }	{ c, A, ρ}	{ k, A, ρ }
{**apv**}	{ A, ρ }	{ A, ρ }	{ k, A, ρ }
{**apf**}	{ c }	{ c }	{ }
{**ocu**}	{k}	{ }	{ }

2.2.2. Processamento incorreto da informação

O problema aqui apresentado tem características que o tornam muito comum na prática da engenharia: a necessidade de reconstituir uma verdade oculta, inacessível diretamente, por meio de informações limitadas e premissas duvidosas. Para tanto, as informações disponíveis devem ser processadas de acordo com a descrição ideal apresentada na Equação (72) de modo a garantir a equivalência com a realidade do processo representado. Este processamento se ampara em três etapas fundamentais: o recorte do conjunto de informações necessárias (Equações 15 e 71), a consolidação do conjunto de informações disponíveis (Equação 51), e o cálculo das variáveis conseqüentes (Equação 14).

$$\mathbf{ZZa}(\mathbf{ms}^{-}),\ \mathbf{ZZm}_0(\mathbf{in} \setminus \mathbf{ms}^{-}) \xrightarrow{\mathcal{T}\mathcal{P}m} \mathbf{ZZm}(\mathbf{in}) \Leftrightarrow \mathbf{ZZ}(\mathbf{in}) \xrightarrow{Tproc} \mathbf{ZZm}(\mathbf{out}) \Leftrightarrow \mathbf{ZZ}(\mathbf{out}) \tag{72}$$

A abordagem adotada no presente texto classificará dentro da denominação de *processamento incorreto da informação* os fatores que coloquem em risco as transformações que permitem, a partir da informação preliminar disponível, reconstituir o análogo conhecido do processo, **ZZm**, como descrito na Equação (72).

As possíveis causas de processamento incorreto são: a projeção distorcida de **ZZ**(**in**) no plano das informações *a priori*, originado pela incapacidade de $\mathcal{P}$m assimilar novas informações devido a informações preliminares incorretas (tipo 1); o uso de uma representação incompleta das variáveis **ZZ** (tipo 2); a suposição incorreta acerca da natureza das relações funcionais ***f*** e ***g*** (tipo 3), que pode estar associada ao emprego de variáveis cujas informações são qualitativamente diferentes daquelas presentes em **ZZ** (tipo 4). Estes tipos são sumarizados formalmente na descrição apresentada abaixo:

Tipo 1) uso de conjunto incorreto de informações: **atr** ≠ **apv** na Equação 67.

Tipo 2) sub-representação das variáveis pertinentes:

$$\mathcal{P}m = \{\mathbf{ZZm} \in \mathrm{R}^{\dim(\mathbf{mod})} \mid \boldsymbol{fm} \wedge \boldsymbol{gm}\}$$

$$\{\ \mathbf{ZZm}\} \ = \ \{\mathbf{ZZ}(\mathbf{mod})\},\ \dim(\mathbf{mod}) < \dim(\mathbf{ZZ}) \tag{73}$$

Tipo 3) erro na estrutura das relações funcionais

$$\boldsymbol{fm} \neq \boldsymbol{f}, \boldsymbol{gm} \neq \boldsymbol{g} \tag{74}$$

Tipo 4) uso de conjunto distinto de variáveis

$$\mathcal{P}m' = \{\mathbf{ZZm} \in \mathsf{R}^{\dim(\mathbf{ZZm})} \mid \boldsymbol{fm} \wedge \boldsymbol{gm}\} \tag{75}$$
$$\exists i \in \mathsf{N} : \{\mathbf{ZZm}(i)\} \notin \{\mathbf{ZZ}\}$$

Deve-se notar que os quatro tipos acima elencados muito provavelmente apresentar-se-ão combinados como causas de falhas no correto encaminhamento das transformações descritas na Equação (72). A presente abordagem procura dar um cunho mais genérico a estes processos, explicitando um foco mais largo que o conceito de *model mismatch* comumente encontrado na literatura [12, 40, 52], e que constitui o processamento incorreto do tipo 3.

2.2.3. Informação corrompida

A apropriação da verdade acerca do processo $\mathcal{P}$ por meio de observações diretas, conforme expresso na Equação (49), é improvável em sistemas reais. Na verdade, cada observação é realizada por meio da incorporação de um instrumento físico à envoltória do processo. Cada incremento de conhecimento a respeito do processo é obtido por meio de comparações com análogos da informação real, mediados por processos também sujeitos, em termos genéricos, a uma estrutura similar à Equação (11). Desta forma, cada medição incorpora o detalhamento da descrição do processo nas vizinhanças do instrumento físico de medição, incluindo os balanços das propriedades conservativas na interface processo/instrumento e o mecanismo de transdução entre grandezas físicas análogas. O processo expandido pela inclusão dos N elementos de medição passa então a ser representado pela Equação (76).

$$\mathcal{P}^{+} = \{\underbrace{[\mathbf{ZZ}^{\mathrm{T}}\ \mathbf{ZZ}_{pm1}^{\mathrm{T}} \ldots \mathbf{ZZ}_{pmN}^{\mathrm{T}}]^{\mathrm{T}}}_{\mathbf{ZZ}^{+}} \in \mathsf{R}^{\dim(\mathbf{ZZ}^{+})} \mid \boldsymbol{f} \wedge \boldsymbol{g} \wedge \boldsymbol{f}_{pm1} \wedge \boldsymbol{g}_{pm1} \wedge \ldots \wedge \boldsymbol{f}_{pmN} \wedge \boldsymbol{g}_{pmN}\} \tag{76}$$

Em termos formais, cada modelo de processo de medição deveria ser incorporado, de forma completa e detalhada, ao processo global, $\mathcal{P}m$. Contudo, a elevada dimensionalidade do problema resultante pode ser um obstáculo à utilização, assim como o grande número de relações funcionais de difícil caracterização, o que pode tornar contraproducente a tentativa de representação explícita e completa do processo de medição. Isto ocorre porque o incremento de informação de cada processo de medição no conjunto **ms** dificilmente compensaria o acréscimo dimensional associado aos conjuntos de informações ocultas, **ocu**, e de hipóteses falsas, **apf**.

É comum supor que a informação obtida pelos processos de observação direta possa ser representada, de forma global, como o mapeamento da Equação (77), que incorpora, aditivamente à informação real, o termo $\boldsymbol{\varepsilon}$, associado à possível corrupção do sinal ao longo do processo de observação (Equação 78). Estas funções, em conjunto com o processo $\mathcal{P}m$, são as versões alternativas e disponíveis da realidade descrita pela Equação (76).

Caso os processos e instrumentos de medição sejam de tal forma que coletem diretamente a informação pertinente, sem fazer uso intermediário de análogos físicos, e se a medição não for influenciada por quaisquer propriedades do instrumento, quer de constituição, de geometria ou de processamento da informação, o processo $\mathcal{P}^+$ se reduz a $\mathcal{P}$. Neste caso, o processo de medição da Equação (78) dá conta apenas de influências cuja evolução não é completamente definida pelas condições iniciais, caracterizando variáveis aleatórias que são realizações de eventos circunscritos por funções distribuição de probabilidade (Equação 79). Porém, em caso contrário, $\boldsymbol{\varepsilon}$ expressará a acomodação de todos os elementos pertencentes a $\mathcal{P}^+$ e não explicitamente contidos em $\mathcal{P}$ ou em $\mathcal{P}m$, não necessariamente resultando em uma variável estocástica *stricto sensu*, embora possa aparentar esta propriedade se o número destes elementos for suficientemente grande, variar em elevada freqüência e em escalas similares.

$$\begin{aligned} &Tmed : \mathbb{R}^{\dim(\mathbf{ms})} \rightarrow \mathbb{R}^{\dim(\mathbf{ms})} \\ &\mathbf{ZZ}(\mathbf{ms}) \mapsto \mathbf{ZZa}(\mathbf{ms}) \end{aligned} \tag{77}$$

$$\mathbf{ZZa} = \mathbf{x}^{\mathrm{T}}(\mathbf{ZZ} + \boldsymbol{\varepsilon})$$
$$se\ i \in \mathbf{ms},\ \mathbf{x}(i)=1,\ P(\mathbf{ZZa}(i) \neq \mathbf{ZZ}(i)) > 0 \qquad (78)$$
$$se\ i \notin \mathbf{ms},\ \mathbf{x}(i)=0$$

$$\boldsymbol{\varepsilon} \sim \psi(\boldsymbol{\varepsilon}) \ \rightarrow\ \mathbf{ZZa} \sim \psi(\mathbf{ZZa}) \qquad (79)$$

$$\mu(\boldsymbol{\varepsilon}(i)) = 0 \ \rightarrow\ \mu(\mathbf{ZZa}(i)) = \mathbf{ZZ}(i)$$

2.3. Adaptação do Modelo

Em função da formulação escolhida para o problema, parte das informações obtidas por observação direta pode não ser apropriada pelo conjunto de variáveis necessárias (via mapeamento $T\wp m$ (Equações 51 e 52)), de modo que $\dim(\mathbf{ms}^{-}) < \dim(\mathbf{ms})$. Neste caso, haverá um conjunto de variáveis, referenciadas pelo vetor de índices **dual** (Equação 80), que faz parte simultaneamente do conjunto de variáveis conseqüentes e das informações obtidas por observação direta. Estas variáveis duais indicam a existência de um excesso de informação originado pela estrutura de configuração do problema, evidenciado pela duplicidade de origens a partir das quais seus valores podem ser assinalados (81).

$$\mathbf{dual} = \mathbf{ms} \setminus \mathbf{ms}^{-} = \mathbf{out} \cap \mathbf{ms} \qquad (80)$$

$$\left.\begin{array}{l} \mathbf{ZZm(in)} \xrightarrow{Tproc} \\ \mathbf{ZZa(ms)} \longrightarrow \end{array}\right| \mathbf{ZZm(dual)} \qquad (81)$$

A informação latente armazenada em **ZZa**(**dual**) pode ser desdobrada de modo a suprir conhecimento útil para tornar mais completa e atualizada a descrição do estado atual de **ZZ**(**in**). Em um cenário ideal, isto implicaria:

- Obter informações a respeito das mudanças em variáveis necessárias que não podem ser observadas diretamente. O conjunto completo das variáveis em **ZZ** que se enquadra nesta descrição é referenciado pelo vetor de índices **Est** (Equação 82):

$$\mathbf{Est} = (\mathbf{in} \cap \mathbf{var}) \setminus \mathbf{ms} \tag{82}$$

- Corrigir os efeitos da corrupção do sinal de variáveis observadas diretamente, referenciadas pelos índices **crp** (Equação 83). O conjunto completo das variáveis em **ZZ** que devem ser adaptados de modo a reduzir estes efeitos é referenciado, idealmente, pelo vetor de índices **Rec** (Equação 84).

$$\mathbf{crp} = \{\mathbf{x} \mid \mathbf{crp} \subseteq \mathbf{ms}\,,\ P((\boldsymbol{\varepsilon}(\mathbf{x}) \neq \mathbf{0}) > \mathbf{0})\} \tag{83}$$

$$\mathbf{Rec} = \mathbf{crp} \tag{84}$$

Desta forma, a incompletude das informações obtidas por observação direta (violação da condição expressa em (60)) e a corrupção das informações do processo de medição definem o conjunto dos elementos cuja observação indireta é requerida. Seguindo as necessidades acima descritas, este conjunto é dado por **Upd** (85).

$$\mathbf{Upd} = \mathbf{Rec} \cup \mathbf{Est} \tag{85}$$

Os tipos de informação acima descritos, requeridos para a completa reconstituição da verdade **ZZ,** dão origem a distinções de nomenclatura classicamente [53] adotadas na literatura: o problema de estimação busca propor valores aos elementos referenciados por **est**, assinalando-os por meio das funções de atribuição, de acordo com a Equação (86); o problema de reconciliação trata de propor valores aos elementos referenciados por **rec**, assinalando-os por meio das funções de medição, de acordo com a Equação (87).

Note-se que o problema de adaptação, que unifica ambos os casos, consiste em encontrar os valores das correções aditivas **Θ** para os elementos **upd** (Equação 88).

$$\begin{aligned} &f_{atr} : \mathbf{ZZm}_j(\mathbf{est}) = \mathbf{ZZm}_0(\mathbf{est}) + \boldsymbol{\Theta}(\mathbf{est}) \\ &\mathbf{est} \cap \mathbf{ms} = \phi \end{aligned} \tag{86}$$

$$f_{med}:\ \mathbf{ZZm}_j(\mathbf{rec}) = \mathbf{ZZa}_j(\mathbf{rec}) + \boldsymbol{\Theta}(\mathbf{rec})$$
$$\mathbf{rec} \subset \mathbf{ms} \tag{87}$$

$$\mathbf{upd} = \mathbf{rec} \cup \mathbf{est} \tag{88}$$

Em termos genéricos, o problema consistiria em, partindo-se do conjunto de informações aparentes acumuladas ao longo da evolução do processo, Q_a (89), e aproveitando a informação latente contida nas variáveis duais, gerar o conjunto Q_a^+ (90), que contém a "melhor" representação observável possível da história do processo até o instante mais recente, j.:

$$Q_a = \left\{ \mathbf{ZZm}_{0..j-1},\ \mathbf{ZZa}_j(\mathbf{ms}),\ \boldsymbol{fm},\ \boldsymbol{gm} \right\} \tag{89}$$

$$Q_a^+ = \left\{ Q_a,\ \mathbf{ZZm}_j \right\} \tag{90}$$

Qualitativamente, os requisitos para a assimilação completa da variabilidade apresentada por **ZZ** são: que o conjunto das variáveis atualizáveis seja idêntico ao conjunto cuja atualização é requerida para suprir as deficiências (Equações 82-88) de completude e corrupção das informações, como previsto na Equação (91); que o conjunto de informações *a priori* verdadeiras descreva de forma completa o conjunto de variáveis atribuídas, como previsto na Equação (92).

$$\mathbf{upd} = \mathbf{Upd} \tag{91}$$

$$\mathbf{atr} = \mathbf{apv} \tag{92}$$

O mapeamento *Tadap* (Equação 93), produz o conjunto de informações indiretas, *Ioi*, composto pela união das variáveis consequentes (**out**) e atualizáveis (**upd**). Sob a validade das Equações (91-92), existem as bases qualitativas para a reconstrução da verdade **ZZm** = **ZZ** com o auxílio dos modificadores **Θ**.

$$Tadap : \mathrm{R}^{\dim(\mathbf{ms})+\dim(\mathbf{atr})} \rightarrow \mathrm{R}^{\dim(\mathbf{out})+\dim(\mathbf{upd})}$$

$$\left(\underbrace{\mathbf{ZZm_0(atr)}}_{Iap}, \underbrace{\mathbf{ZZa_j(ms)}}_{Iod} \right) \mapsto \underbrace{\left[\mathbf{ZZm(out)}^{\mathrm{T}} \quad \mathbf{ZZm(upd)}^{\mathrm{T}} \right]^{\mathrm{T}}}_{Ioi} \tag{93}$$

2.4. Otimização

O objetivo final do RTO consiste em encontrar os valores das variáveis correspondentes ao subconjunto de índices **df** em **ZZm,** que indicam os graus de liberdade disponíveis para a melhoria da função de desempenho econômica, L. As variáveis referenciadas por **df** constituem as variáveis de decisão do RTO e devem, obrigatoriamente: 1) fazer parte do elenco escolhido de variáveis necessárias (Equação 94); 2) ser diretamente observáveis, e 3) fazer parte do conjunto de entradas do processo (**II**, na Equação 25), de acordo com a Equação (95).

$$\mathbf{df} \subseteq \mathbf{in} \tag{94}$$

$$\mathbf{df} \subseteq (\mathbf{ms} \cap \mathbf{II}) \tag{95}$$

Exemplo 5 – Escolha do conjunto de variáveis de decisão

Sendo o processo apresentado nos exemplos 1 e 2, cujo conjunto de entradas e variáveis medidas são dados por:

$$\{\mathbf{II}\} = \{Fe, x_{1,e}, c, k, Fs, A, \rho\}$$

$$\{\mathbf{ms}\} = \{Fe, x_{1,e}, h, F_s\}$$

Se as condições necessárias para a resolução do modelo matemático foram definidas como:

in = {Fe,$x_{1,e}$, c, k, A, ρ}

Então, as variáveis de decisão devem ser estar contidas no subconjunto abaixo:

$\mathbf{df} \subseteq (\mathbf{in} \cap \mathbf{ms} \cap \mathbf{II})$

$\{\mathbf{df}\} \subseteq \{Fe, x_{1,e}\}$

A escolha dos valores mais convenientes para as variáveis de decisão corresponde ao problema de otimização não linear descrito em (96), vinculado ao conjunto $\mathcal{M}$ (97), que contém o conjunto de métodos e algoritmos, alg, sob a parametrização $\boldsymbol{\theta}_{alg}$, cujo produto final se materializa na formulação dos valores das variáveis de decisão previstos no instante j e implementados no instante subseqüente, $\mathbf{ZZm}_{j+1}(\mathbf{df})\big|_j$.

$$\mathbf{ZZm}_{j+1}(\mathbf{df})\big|_j = \underset{\mathbf{ZZm}_j(\mathbf{df})}{\arg\min}\left(\mathrm{L}\left(\mathbf{ZZm}_j\right)\right)\Big|_{\mathcal{M}}$$

$$s.a \qquad (96)$$

$$\mathrm{Pm}: \begin{cases} \boldsymbol{fm}_{sis} \\ \mathbf{gm} \end{cases}$$

onde

$$\mathcal{M} = \left\{ alg, \boldsymbol{\theta}_{alg} \right\} \qquad (97)$$

O problema formulado em (96) pode ser representado pelo mapeamento *Totm*, descrito em (98).

$$Totm: \mathrm{R}^{\dim(\mathbf{in})} \rightarrow \mathrm{R}^{\dim(\mathbf{df})}$$
$$\left(\mathbf{ZZm}_j(\mathbf{in}) \mapsto \mathbf{ZZm}_{j+1}(\mathbf{df})\big|_j\right)\Big|_{\mathcal{M}.\mathrm{Pm}} \qquad (98)$$

O estado subseqüente do processo sujeito à implementação das variáveis de decisão e à manutenção das demais condições de contorno, $\mathbf{ZZm}_{j+1}\big|_j$, é dado pelo mapeamento *Tprev*, que concatena diversas informações, como descrito nas Equações (99,100).

$$Tprev : \mathrm{R}^{\dim(\mathbf{ZZm})} \rightarrow \mathrm{R}^{\dim(\mathbf{ZZm})}$$
$$\left(\mathbf{ZZm}_j \mapsto \mathbf{ZZm}_{j+1}\Big|_j\right)\Big|_{\mathcal{m}.\mathrm{P}\mathrm{m}} \tag{99}$$

$$\mathbf{ZZm}_{j+1}\Big|_j = Tprev(\mathbf{ZZm}_j) \Leftrightarrow \begin{cases} \mathbf{ZZm}_{j+1}(\mathbf{in} \setminus \mathbf{df})\Big|_j = \mathbf{ZZm}_j(\mathbf{in} \setminus \mathbf{df}) \\ \mathbf{ZZm}_{j+1}(\mathbf{df})\Big|_j = Totm\left(\mathbf{ZZm}_j(\mathbf{in})\right) \\ \mathbf{ZZm}_{j+1}(\mathbf{out})\Big|_j = Tproc\left(\mathbf{ZZm}_{j+1}(\mathbf{in})\Big|_j\right) \end{cases} \tag{100}$$

É importante ressaltar que regras mais genéricas podem ser usadas para implementar os novos valores de decisão, levando em conta informações dos valores medidos e estimados ao longo de um horizonte pregresso de variações. Este procedimento é comum quando se vincula a solução atual às soluções dos ciclos anteriores, conferindo um caráter de filtragem à função *fnc* em (101):

$$\mathbf{ZZm}_{j+1}(\mathbf{df})\Big|_j = fnc\left(Totm\left(\mathbf{ZZm}_j(\mathbf{in})\right), \mathbf{ZZm}_{j-h\ldots j}, \mathbf{ZZa}_{j-h\ldots j}\right),\ h \in \{0,..,j\} \tag{101}$$

O valor da função econômica previsto para j+1 com base nas decisões e condições de contorno em j será dado pela Equação (102):

$$Lm_{j+1}\Big|_j = \mathrm{L}\left(\mathbf{ZZm}_{j+1}\right)\Big|_j = \mathrm{L}\left(Tprev(ZZm_j)\right) \tag{102}$$

2.5. Representação reduzida do problema de RTO: O caso estacionário

É comum a implementação, na prática industrial, de sistemas de otimização cujo foco seja tão somente a predição de estados estacionários nos quais a métrica de desempenho pertinente ao processo seja levada a um ponto extremo. Neste caso, em teoria, existem duas possibilidades de implementação:

1) Não são impostas restrições funcionais sobre a representação da estrutura do processo. Deste modo, o conjunto de informações que

descreve o processo e o conjunto através do qual ele é representado podem, potencialmente, serem os mesmos, {**ZZ**} = {**ZZm**}, assim como ocorreria com as relações funcionais. Neste caso, ainda que o problema de otimização de desempenho esteja restrito à estacionariedade e o espaço de busca esteja reduzido a $R^{\dim(\mathbf{Z})}$ (onde **Z** é descrito pela Equação (30)), a representação completa, dinâmica, estaria disponível para os demais problemas acessórios, como a adaptação do modelo e o suporte a decisões concernentes ao tratamento das informações medidas.

2) Todo o conhecimento e representação estão reduzidos às condições de estacionariedade. Neste caso, todas as instâncias do RTO estarão sujeitas à redução das informações ao conjunto de variáveis **Z** e do conjunto de relações funcionais algébricas.

Este último caso representa o cenário comumente encontrado nas aplicações comerciais em uso na indústria para problemas de larga escala. Isto significa que, ainda que o processo produza informações baseadas no modelo previsto pelas Equações (103), todas as tomadas de decisão baseiam-se na suposição de que seja possível a representação reduzida proposta pelas Equações (104), características da estacionariedade.

Modelo do Processo dinâmico:

$$f_{sis}:\begin{cases} \boldsymbol{f}: \quad \boldsymbol{f}(\mathbf{ZZm})=\mathbf{0} \\ \boldsymbol{f}_{\mathbf{atr}}: \ \mathbf{ZZm}_j(\mathbf{atr}) = \mathbf{a} \\ \boldsymbol{f}_{med}: \mathbf{ZZm}_j(\mathbf{ms}^{-}) = \mathbf{ZZa}_j(\mathbf{ms}^{-}) \end{cases} \tag{103}$$

Modelo do processo estático:

$$f_{sis,E}:\begin{cases} \boldsymbol{f}_E: \quad \boldsymbol{f}(\mathbf{Zm})=\mathbf{0} \\ \boldsymbol{f}_{\mathbf{atr},E}: \mathbf{Zm}_j(\mathbf{atr}_1) = \mathbf{a} \\ \boldsymbol{f}_{med,E}: \mathbf{Zm}_j(\mathbf{ms}_1^{-}) = \mathbf{Za}_j(\mathbf{ms}_1^{-}) \end{cases} \tag{104}$$

Exemplo 6 – Representação reduzida: O caso estacionário

Partindo do modelo apresentado no exemplo 1, temos a seguinte representação:

$$\{\mathbf{ZZ}\} = \left\{ Fe, x_{1,e}, c, k, A, \rho, Fs, h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\}$$

$$\boldsymbol{f}: \begin{cases} \boldsymbol{f}_{alg}: \{ Fs - c\sqrt{h} = 0 \\ \boldsymbol{f}_{dif}: \begin{cases} Fe - Fs - A\rho\dfrac{dh}{dt} = 0 \\ Fe\,x_{1,e} - Fs\,x_1 - kx_1Ah - A\rho\dfrac{dhx_1}{dt} = 0 \end{cases} \end{cases}$$

Informações pertinentes à representação estacionária:

$$\{\mathbf{Zm}\} = \{\mathbf{Z}\} = \{\mathbf{I}, \mathbf{O}\} = \{ Fe, x_{1,e}, c, k, Fs, h, x_1 \}$$

$\{\mathbf{Zm} \setminus \mathbf{ZZ}\} = \{\boldsymbol{\tau}, \mathbf{dO}\} = \{ A, \rho, dh/dt, dhx_1/dt \}$, *variáveis exclusivas do modo dinâmico*

$$\boldsymbol{f}_E: \begin{cases} Fs - c\sqrt{h} = 0 \\ Fe - Fs = 0 \\ Fe\,x_{1,e} - Fs\,x_1 - kx_1Ah = 0 \end{cases}$$

É importante ressaltar que, caso $\mathbf{Zm} \setminus \mathbf{ZZ} \neq \mathbf{0}$ as relações de constância e nulidade deste conjunto que estão implicitamente contidas na representação reduzida não serão válidas, o que implica na ocorrência de processamento incorreto da informação de tipo 2, de acordo com o descrito na Seção 2.2.2.

2.6. Decisões estruturais primárias do RTO

O desempenho do sistema de otimização em tempo real será diretamente dependente das diversas escolhas realizadas para a concatenação de sua estrutura. Esta seção tem por finalidade enumerar estas escolhas, em face de uma rede de instrumentação previamente determinada.

2.6.1. Escolha do conjunto de variáveis necessárias

O conjunto de variáveis necessárias deve ter seus elementos escolhidos dentre os componentes de **ZZ** e possuir dimensão igual à diferença entre o número total de variáveis e a soma do número de relações funcionais do modelo matemático que descreve o processo, ***f***. Deve-se também deduzir a dimensão do conjunto **dO**, que designa as derivadas das variáveis originais, não incluídas no rol das variáveis necessárias. O resultado é expresso pelo número de combinações equivalente ao representado nas Equações (105,106).

Conjunto de escolhas: dim(**ZZm**)

Tamanho dos grupos: $\dim(\mathbf{in}) = \dim(\mathbf{ZZm}) - \dim(\boldsymbol{f}) - \dim(\mathbf{dO})$

Quantidade de grupos:

$$\dim(\mathbf{In}) = C_{\dim(\mathbf{ZZm}),\,\dim(\mathbf{in})} \tag{105}$$

$$C_{\dim(\mathbf{ZZm}),\,\dim(\mathbf{in})} = \frac{\dim(\mathbf{ZZm})!}{\left(\dim(\mathbf{ZZm}) - \dim(\boldsymbol{f}) - \dim(\mathbf{dO})\right)!\ \left(\dim(\boldsymbol{f}) + \dim(\mathbf{dO})\right)!} \tag{106}$$

Contudo, em um problema real, alguns dos elementos de **in** podem ser fruto de escolhas predeterminadas. O número total destas escolhas prévias, n_{pre} (107), reflete os seguintes fatos: os dim(**df**) elementos que compõem o conjunto das variáveis de decisão devem fazer parte de **in**, de acordo com (94); um total de n_{pre_out} variáveis podem ter sido previamente escolhidas para fazer parte do conjunto de variáveis consequentes; n_{pre_in} variáveis podem ter sido previamente escolhidos para fazer parte de **in**. Estas escolhas prévias limitam a dimensão de **In** ao total de combinações apresentadas em (108).

$$n_{pre} = \underbrace{\dim(\mathbf{df})}_{n_u} + n_{pre\_in} + n_{pre\_out} \tag{107}$$

$$\dim(\mathbf{In}) = C_{\dim(\mathbf{ZZm})-n_{pre},\,\dim(\mathbf{in})-n_{pre\_in}-n_u} \tag{108}$$

Em virtude do fato de que a restrição genérica de colinearidade (16) deve ser respeitada, o total de escolhas disponíveis, Ne_{in}, para compor o elenco de variáveis necessárias, apresenta o limite superior indicado na Equação (109)

$$Ne_{\mathbf{in}} \leq \frac{(\dim(\mathbf{ZZm}) - n_{pre})!}{(\dim(\mathbf{in}) - n_{pre\_in} - n_u)!\ (\dim(\mathbf{ZZm}) - n_{pre} - (\dim(\mathbf{in}) - n_{pre\_in} - n_u))!} \tag{109}$$

Exemplo 7 – Escolhas do conjunto de variáveis necessárias

Continuando o processo apresentado nos exemplos 1 e 2, expresso pelas variáveis:

$\{\mathbf{ZZm}\} = \left\{ Fe, x_{1,e}, c, k, A, \rho, Fs, h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\}$, $\dim(\mathbf{ZZm}) = 11$

P*odemos supor algumas pré-escolhas ´já feitas pelo usuário:*

O conjunto de variáveis de decisão já foi definido:

$\{\mathbf{df}\} = \{Fe\} \Rightarrow n_u = 1$

A composição da entrada deve fazer parte do conjunto de variáveis necessárias:

$\{\mathbf{in}\} \subseteq \{x_{1,e}\} \Rightarrow n_{pre\_in} = 1$

As variáveis de estado e as derivadas farão parte do conjunto de variáveis consequentes:

$\{\mathbf{out}\} \subseteq \left\{ h, x_1, \frac{dh}{dt}, \frac{dhx_1}{dt} \right\} \Rightarrow n_{pre\_out} = 4$

$n_{pre} = 1 + 1 + 4 = 6$

Tamanho do conjunto de variáveis necessárias:

$\dim(\mathbf{in}) = \dim(\mathbf{ZZm}) - \dim(\boldsymbol{f}) - \dim(\mathbf{dO}) = 11 - 3 - 2 = 6$

O total de escolhas referentes às variáveis necessárias fará parte de um conjunto cujo tamanho será menor que o limite superior:

$$Ne_{\mathbf{in}} \leq \frac{(\dim(\mathbf{ZZm}) - n_{pre})!}{(\dim(\mathbf{in}) - n_{pre\_in} - n_u)!\ (\dim(\mathbf{ZZm}) - n_{pre} - (\dim(\mathbf{in}) - n_{pre\_in} - n_u))!}$$

$$Ne_{\mathbf{in}} \leq \frac{(11-6)!}{(6-1-1)!\ (11-6-(6-1-1))!}$$

O que resulta um máximo de 5 escolhas para as variáveis necessárias, as quais são representadas pelos seguintes conjuntos:

{Fe, $x_{1,e}$, c,k,A,ρ}	{Fe, $x_{1,e}$, c,k, ρ,Fs}	{Fe, $x_{1,e}$, k,A,ρ,Fs}
{Fe, $x_{1,e}$, c,k,A,Fs}	{Fe, $x_{1,e}$, c,A,ρ,Fs}	

O atendimento à restrição de colinearidade da Equação (16) *definirá o conjunto definitivo de escolhas **In**. No presente caso, todas as possibilidades a atendem.*

2.6.2. Escolha do conjunto de variáveis atualizáveis

As escolha do conjunto **upd** deve representar um subconjunto de **ZZm** cuja dimensão não seja superior à quantidade de variáveis diretamente observadas, respeitando o fato de que nem todas as variáveis duais podem ser simultaneamente reconciliadas, como representado em (110). Caso contrário, a solução trivial $\boldsymbol{\theta} = \boldsymbol{\Theta}(\mathbf{rec}) = \mathbf{0}$ seria a resposta do problema de adaptação. Além disto, as variáveis estimadas não devem fazer parte do conjunto de variáveis de saída do modelo.

$$\begin{cases} \mathbf{upd} \subset \{1,2...\dim(\mathbf{ZZm})\} \\ \dim(\mathbf{upd}) \leq \dim(\mathbf{ms}) \\ \dim(\mathbf{dual}) > 0 \Leftrightarrow \dim(\mathbf{out} \cap \mathbf{ms}) > 0 \\ \mathbf{dual} \neq \mathbf{rec} \end{cases} \tag{110}$$

A escolha de cada possível conjunto **upd,** condicionada à escolha de cada **in**, está contida em **Gupd**, conforme explicitado na Equação (111). O número total de escolhas de **upd** dado **in** é resultado da escolha feita no universo de subconjuntos formados pelos índices de **ZZm** de dimensão não maior que o número de medições, e que contenha ao menos um elemento que se refira a uma variável dual determinada pela prévia escolha de **in**, como sumarizado em (113). O conjunto **Rto**, de todas as estruturas possíveis baseadas em escolhas de **in** e **upd**, é expresso nas Equações (114-115).

$$\mathbf{Gupd}(\bullet)|_{\mathbf{In}\{x\}} = \left\{ \mathbf{c} \,|\, \mathbf{c} \subset \{1,2...\dim(\mathbf{ZZm})\},\ \dim(\mathbf{c}) \leq \dim(\mathbf{ms}),\ \mathbf{dual}|_x \neq \mathbf{rec} \right\}$$
$$x \in \{1,..,Ne_{in}\} \tag{111}$$

$$\mathbf{dual}|_x = \mathbf{ms} \setminus \mathbf{In}(x) \tag{112}$$

$$Ne_{upd|In(x)} = \dim(\mathbf{Gupd}|_{\mathbf{In}\{x\}}) \tag{113}$$

$$\mathbf{Rto}(x,\ y) = \left\{ \mathbf{In}(x),\ \mathbf{Gupd}(y)|_{\mathbf{In}(x)} \right\} = \left\{ \mathbf{in}, \mathbf{upd} \right\}_{|y|x} \tag{114}$$

$$\dim(\mathbf{Rto}) = \sum_{x=1}^{Ne_{in}} Ne_{upd|In(x)} \tag{115}$$

2.6.3. Escolha dos elementos da função objetivo de adaptação do modelo

A introdução dos modificadores $\boldsymbol{\Theta}$ no conjunto de variáveis acresce dim(**upd**) novos elementos a **ZZm**, criando o conjunto $\mathbf{ZZm}^{+}$ (116), alterando os graus de liberdade do sistema resultante (117).

$$\mathbf{ZZm}^{+} = [\mathbf{ZZm}^{T} \;\; \boldsymbol{\Theta}(\mathbf{upd})^{T}]^{T}, \;\; \dim(\mathbf{ZZm}^{+}) = \dim(\mathbf{ZZm}) + \dim(\mathbf{upd}) \tag{116}$$

$$GL = \dim(\mathbf{ZZm}^{+}) - \left(\dim(\boldsymbol{f}) + \dim(\boldsymbol{f}_{atr}) + \dim(\boldsymbol{f}_{med})\right) = \dim(\mathbf{upd}) \tag{117}$$

Estes graus de liberdade são consumidos pelo procedimento de adaptação, que faz uso do excesso de informação contido nas variáveis duais. Este procedimento desemboca em um problema de otimização cuja função objetivo é uma métrica de proximidade, *F*, entre as observações diretas e as informações produzidas pelo filtro de observação dado pelas relações $\{\boldsymbol{f}_{sis}, \boldsymbol{gm}\}$, como mostrado na Equação (118).

$$\begin{aligned} &\boldsymbol{\Theta}_{j}(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_{j}(\mathbf{upd})}{\arg\min}\left(F\left(\underbrace{\mathbf{ZZm}(\mathbf{dual})}_{\mathbf{ZZ}_{\text{modelo}}}, \underbrace{\mathbf{ZZa}(\mathbf{dual}), \mathbf{ZZa}(\mathbf{rec})}_{\mathbf{ZZ}_{\text{obs}}} \right) \right) \\ &s.a \\ &\boldsymbol{f}_{sis} \\ &\boldsymbol{gm} \end{aligned} \tag{118}$$

Por motivos relacionados principalmente à identificabilidade dos parâmetros, é comum que nem todas as variáveis duais e reconciliáveis integrem a função objetivo (118), mas apenas aquelas referenciadas pelos índices **obj**, fazendo com que o problema de otimização tome a forma apresentada na Equação (119).

Conforme apresentado na Equação (120), o conjunto de elementos pertencentes à função objetivo deve preencher alguns requisitos: 1) estar contido nas variáveis medidas, 2) não ser menor do que o conjunto de variáveis atualizadas, $\dim(\mathbf{obj}) \geq \dim(\mathbf{upd})$, e conter ao menos uma variável dual que não seja sujeita a reconciliação,

$\mathbf{obj} \cap (\mathbf{dual} \setminus \mathbf{rec}) \neq \phi$, por motivos similares aos apresentados aos formulados quando da apresentação da Equação (110).

$$\mathbf{\Theta}_j(\mathbf{upd}) = \underset{\mathbf{\Theta}_j(\mathbf{upd})}{\arg\min}\left(F\left(\underbrace{\mathbf{ZZm}(\mathbf{obj})}_{\mathbf{ZZ}_{\text{modelo}}}, \underbrace{\mathbf{ZZa}(\mathbf{obj})}_{\mathbf{ZZ}_{\text{obs}}} \right) \right)$$
$$s.a \quad \boldsymbol{f}_{sis} \quad \boldsymbol{gm} \tag{119}$$

$$\mathbf{obj} = \left\{ \mathbf{x} \mid \mathbf{x} \subseteq \mathbf{ms}, \dim(\mathbf{x}) \geq \dim(\mathbf{upd}), \mathbf{x} \subseteq (\mathbf{dual} \cup \mathbf{rec}), \mathbf{x} \cap (\mathbf{dual} \setminus \mathbf{rec}) \neq \phi \right\} \tag{120}$$

O número de escolhas dos elementos do conjunto **obj** condicionadas às escolhas de **upd** e **in** é expresso pelo conjunto **Gobj**, definido de acordo com as Equações (121-123), que formaliza as condições descritas em (120) no contexto das escolhas prévias. Deste modo, o número de diferentes escolhas de **obj** condicionado a dada escolha combinada de **in** e **upd** é descrito de acordo com a Equação (124).

$$\mathbf{Gobj}(\bullet)\big|_{\mathbf{Gupd}(y)|\mathbf{In}(x)} = \left\{ \mathbf{c} \;\middle|\; \begin{cases} \dim(\mathbf{c}) \geq \dim(\mathbf{Gupd}(y) \mid \mathbf{In}(x)) \\ \mathbf{c} \subseteq \left(\mathbf{dual}\big|_x \cup \mathbf{rec}\big|_{x,y} \right) \\ \mathbf{c} \cap \mathbf{est}\big|_{x,y} \neq \phi \end{cases} \right\} \tag{121}$$

$$\mathbf{rec}\big|_{x,y} = \mathbf{Gupd}(y)\big|_{\mathbf{In}(x)} \cap \mathbf{ms} \tag{122}$$

$$\mathbf{est}\big|_{x,y} = \mathbf{Gupd}(y)\big|_{\mathbf{In}(x)} \setminus \mathbf{ms} \tag{123}$$

$$Ne_{obj|Gupd(y)|In(x)} = \dim\left(\mathbf{Gobj}\big|_{\mathbf{Gupd}(y)|\mathbf{In}(x)} \right) \tag{124}$$

A estrutura definida pelas escolhas apresentadas é definida de acordo com o conjunto **Rto** (125), cujo número total de elementos é expresso por Ne_{str}, de acordo com a Equação (126), representando o número total de estruturas que podem ser configuradas para dado problema.

$$\mathbf{Rto}(x,y,z) = \left\{\mathbf{In}(x), \mathbf{Gupd}(y)\,|_{\mathbf{In}(x)}, \mathbf{Gobj}|_{\mathbf{Gupd}(y)|\mathbf{In}(x)}\right\} = \left\{\mathbf{in}, \mathbf{upd}, \mathbf{obj}\right\}_{|z|y|x} \tag{125}$$

$$Ne_{str} = \sum_{x=1}^{Ne_{in}} \sum_{y=1}^{Ne_{upd|In(x)}} Ne_{obj|Gupd(y)|In(x)} \tag{126}$$

3. Aspectos da implementação de otimizadores em tempo real

3.1. Estratégia de solução do problema de RTO

Em plantas industriais, a função de desempenho L na maior parte das vezes tem como núcleo o lucro financeiro, ou seja, a diferença entre receitas e despesas operacionais derivadas de cada escolha de **u** = **ZZ**(**df**) na Equação (96). O modo matematicamente mais simples, embora operacionalmente mais trabalhoso, de formular os valores **u** que conduzem ao máximo valor de L é através do método de busca direta [7,8,9]. Este método prevê a realização de vários experimentos na planta, modificando-se os valores de **u**, de modo a que se identifique o conjunto que conduza à melhor direção do gradiente da função de desempenho. Contudo, o elevado número de passos para alcançar o ótimo compromete o ganho econômico [10]. Exceto para plantas com poucos graus de liberdade e rápida resposta, há pouco apelo para o uso contemporâneo deste método.

Se o problema de otimização não for resolvido por busca direta, outras modificações devem ser feitas em sua formulação. A partição proposta para as variáveis, visando a lidar com condições menos idealizadas, prevê que, conhecendo-se as condições iniciais, as sucessivas implementações do vetor **u** induzirão as correspondentes trajetórias ótimas das respostas da planta, **OO**(t), e do lucro operacional, L(t). Esta é uma abordagem conveniente desde que seja factível obter um modelo dinâmico fidedigno cujo controlador seja facilmente sintonizável [11]. Além disto, a complexidade de $\{\boldsymbol{f}(\mathbf{ZZ})=\mathbf{0}, \boldsymbol{g}(\mathbf{ZZ}) \leq \mathbf{0}\}$ não deve ser tal que impeça a resolução computacional do problema em tempo menor que o período de ocorrência das perturbações**.**

Se estas condições não forem satisfeitas pode ser mais conveniente delegar a tarefa de controlar as trajetórias **u**(t) e **O**(t) a outros sistemas, restringindo o escopo da otimização à designação dos pontos finais dos percursos, quando o estado estacionário é atingido. Isto implica adequar a questão proposta à resolução da classe de problemas em que **dO**=**0** (Equação 28) e, por consequência, em que **ZZ**=**Z**. Desta forma, o problema se torna:

$$u_{j+1}\Big|_j = \mathbf{Z}_{j+1}(\mathbf{df})\Big|_j = \arg\min_{\mathbf{Z}_j(\mathbf{df})}\left(\mathrm{L}\left(\mathbf{Z}_j\right)\right)\Big|_{\mathcal{M}}$$

$$s.a \qquad (127)$$

$$\begin{cases} \boldsymbol{f}_{a\lg} \\ \mathbf{g} \end{cases}$$

Prosseguindo no esforço de formular o conjunto de decisões estruturais de um sistema de otimização em tempo real, o conjunto de escolhas apresentado na Equação (125) pode ser expandido de modo a incluir os graus de liberdade disponibilizados ao otimizador da função econômica, **df**, bem como o método de otimização $\mathcal{M}$ e sua parametrização, associados ao símbolo $\mathcal{M}_{\theta_o}$. Este conjunto de escolhas é apresentado na Equação (128).

$$\mathbf{Rto} \rightarrow \left\{\mathbf{in}, \mathbf{upd}, \mathbf{obj}, \mathbf{df}, \mathcal{M}_{\theta_o}\right\} \qquad (128)$$

Deve-se ressaltar que alguns graus de liberdade associados aos elementos de **in** ficam sob comando dos sistemas que controlam as trajetórias ou estão sob ação direta da ação humana, não sendo escolhas possíveis para **df**.

A formulação do problema de otimização expressa em (96) e em (127) diz respeito, em termos reais, à manutenção das propriedades conservativas e a considerações do mundo real, como as faixas de valores admissíveis por questões de segurança operacional, metalurgia, operação de longo prazo, limites físicos de equipamentos e qualidade de produtos e emissões.

É importante ressaltar a inevitável perda de desempenho da função econômica advinda do fato de a formulação proposta em (127) abstrair a passagem do tempo em face da possível ocorrência posterior das perturbações indicadas pelos índices **d**. Estas perturbações referem-se às variáveis sujeitas à variação temporal (**var**), que fazem parte do conjunto **in** das variáveis necessárias à resolução do modelo e que não são variáveis de decisão do processo de otimização, conforme expresso na Equação (129). Este conjunto **d** pode ser subdividido em um conjunto de índices que referenciam as perturbações medidas, **dms** (Equação (130)) e em outro, que referencia perturbações não medidas, **dum** (Equação (131)).

$$\mathbf{d} = (\mathbf{in} \cap \mathbf{var}) \backslash \mathbf{df} \tag{129}$$

$$\mathbf{dms} = (\mathbf{d} \cap \mathbf{ms}) \tag{130}$$

$$\mathbf{dum} = \mathbf{d} \backslash \mathbf{dms} \tag{131}$$

Suponha-se que ocorram mudanças em valores medidos pertencentes ao vetor de perturbações, **Z**(**d**): Haverá dois intervalos de tempo ao longo do qual a planta funcionará fora de suas condições ótimas. O primeiro está associado ao esforço das malhas de controle para reagir à variação induzida pelas perturbações e manter a planta operando nas condições de referência prévias às perturbações. A perda de desempenho de L é causada pelas trajetórias não otimizadas das variáveis manipuladas pelos controladores e pelo fato de a planta ser conduzida a uma condição operacional comprometida com uma optimalidade tornada obsoleta pela presença das perturbações**.** O segundo intervalo de tempo está associado ao esforço das malhas de controle para, após novo ciclo de otimização, conduzir a planta ao novo estado estacionário, referido à nova condição operacional.

3.1.1. Otimização em duas etapas

O problema de otimização, conforme a formulação (127), admite implicitamente as seguintes premissas:

- o comportamento do processo é estruturalmente descrito de forma perfeita pela representação $\mathcal{P}m$ (Equação (45)).

- As perturbações não medidas, **Z**(**dum**), não variam ao longo da história do processo, ou seja, **dum** $\subseteq$ **fix**.

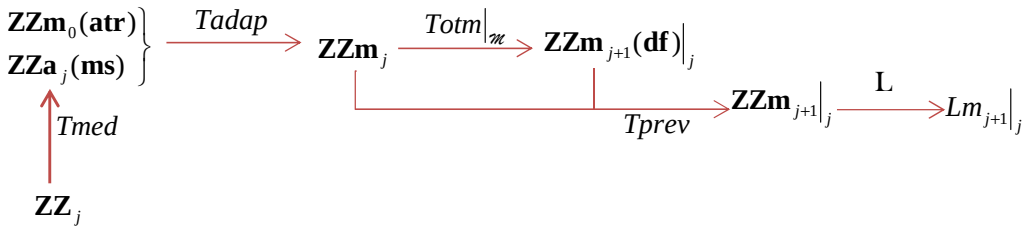
Em casos reais, infelizmente, não apenas **Z**(**dum**) é modificado ao longo do tempo, mas também nosso conhecimento sobre o estado da planta, obtido por intermédio das medições **Z**(**ms**), não está inequivocamente vinculado à realidade. A informação fisicamente contida na variável de processo é corrompida em termos de magnitude e de

dinâmica ao longo dos processos de transdução, transmissão, discretização e armazenamento do sinal. Alguns dos elementos de **Z**(**d)** expressam variações de curto prazo, enquanto outros representam variações graduais no desempenho de equipamentos, tais como a formação de depósitos sólidos em superfícies de troca térmica, variação da eficiência de catalisadores, deterioração física de equipamentos, dentre outras.

Em vista destes fatos, torna-se necessário prever mecanismos de adaptação dos modelos de comportamento que permitam que os valores de **u** formulados a partir de (127) correspondam ao efetivo valor ótimo da função de desempenho. Tradicionalmente, dependendo da classe de variáveis à qual a adaptação se destina, o problema costuma ser nomeado como problema de estimação de parâmetros ou de reconciliação de dados, respectivamente associado às variáveis **est** (Equação 86) e às variáveis **rec** (Equação 87). Embora tal distinção de nomenclatura chegue a acarretar o emprego de procedimentos distintos para sua solução, eventualmente tratada de forma seqüencial ou simultânea [54-57], ambos os problemas possuem a mesma natureza e serão tratados de forma unificada no presente trabalho, como ocorre de forma mais comum na literatura [10,15-26], à parte sua designação por conjuntos de índices distintos.

A adaptação do modelo costuma ser baseada na moldura conceitual apresentada na Seção 2.3, apoiando-se, mais especificamente, no mapeamento previsto na Equação (93). Este mapeamento comumente apresenta-se como resultado de um problema de otimização que espelha a maximização de uma função de verossimilhança [51], produzindo resultados que sejam os mais prováveis dada a natureza das informações disponíveis por observação direta e *a priori*, respectivamente designadas por Iod e Iap.

O mecanismo contido em um ciclo de RTO em malha aberta está esquematizado em (132), e baseia-se em duas camadas de otimização sucessivas, respectivamente associadas aos mapeamentos *Tadap* (Equação 93) e *Totm* (Equação 98). Em dado instante j, o RTO se apropria das informações disponíveis e gera quatro produtos principais: uma versão atualizada do processo, $\mathbf{ZZm}_{j}$; uma proposta de manipulação dos graus de liberdade disponíveis a ser aplicada no passo seguinte, $\mathbf{ZZm}_{j+1}(\mathbf{df})|_{j}$; a previsão do estado futuro, $\mathbf{ZZm}_{j+1}|_{j}$, via *Tprev* (Equações 99-100), e a função de desempenho a ser atingida em consequência da implementação proposta, $\mathrm{Lm}_{j+1}|_{j}$ (Equação 102).



(132)

A Figura 3 apresenta o fluxo de informação através do RTO em duas camadas, onde podem ser vistas as consequências da implementação das decisões no instante j, $\mathbf{Z}_j(\mathbf{df})$, ficando claro que a abordagem de otimização em duas camadas descrita esquematicamente na Equação (132) possui o apelo da simplicidade da forma. O modo como incorpora as informações oriundas do processo real para melhorar a qualidade das previsões do modelo está pretensamente amparado em fundamentos estatísticos consolidados. Resta saber, porém, se estas qualidades se conjugam coletivamente de modo a garantir o resultado ótimo, qual seja, o de manter a operação da planta no estado de melhor desempenho sob a ótica da métrica selecionada.

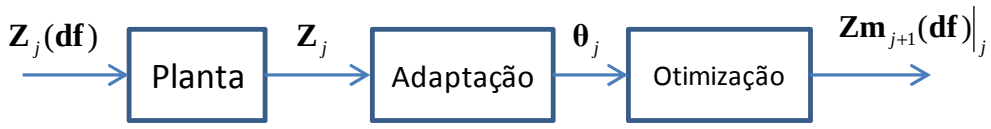


Figura 3 – Otimização em duas etapas

À medida que o processo evolui ao longo de sucessivas intervenções do RTO, sua trajetória apresentará uma variabilidade que é condicionada, por um lado, pelo cenário externo, $\mathbf{Z}(\mathbf{d})$, e, por outro, pelo modo como o RTO reage a elas. Esta variabilidade pode ser expressa como apresentado na Equação (133), para **Z**, e na Equação (134), para L. O caráter genérico de $\lambda_\bullet$ serve para indicar a liberdade de formulação da métrica de variabilidade υ, de acordo com a conveniência de seu uso.

Em condições ideais de completude, processamento e fidelidade das informações disponíveis, a constância do cenário de operação produzirá valores nulos da variabilidade medida (Equação 135), qualquer que seja o modo pelo qual ela for definida.

$$\lambda_Z(i) = \upsilon\{\mathbf{Z}_{j+1}(i),...,\mathbf{Z}_{j+k+1}(i)\}, \quad i \in \{1,2,...,\dim(\mathbf{Z})\} \tag{133}$$

$$\lambda_L = \upsilon\{\mathrm{L}_{j+1}, ..., \mathrm{L}_{j+k+1}\} \tag{134}$$

$$\mathbf{Z}(\mathbf{d})=0 \Rightarrow \boldsymbol{\lambda}_Z = \mathbf{0},\ \lambda_L = 0 \tag{135}$$

É importante ressaltar o caráter inacessível da informação contida nas medidas de variabilidade contidas nas Equações (133-134). Tal informação encontra-se disponível por meio de seus análogos, tais como observados pelo RTO. Deste modo, as variabilidades de observação e de predição, que são a fonte de análise de desempenho, são acessíveis, respectivamente, por meio das Equações (136-137) e (138-139).

$$\boldsymbol{\lambda}\mathbf{m}_Z(i) = \upsilon\{\mathbf{Zm}_{j+1}(i), ..., \mathbf{Zm}_{j+k+1}(i)\},\quad i \in \{1,2,...,\dim(\mathbf{Zm})\} \tag{136}$$

$$\lambda m_L = \upsilon\{\mathrm{Lm}_{j+1}, ..., \mathrm{Lm}_{j+1+k}\} \tag{137}$$

$$\boldsymbol{\lambda}\mathbf{m}_Z(i)\big|_{+1} = \upsilon\{\mathbf{Zm}_{j+1}(i)\big|_j, ..., \mathbf{Zm}_{j+k+1}(i)\big|_{j+k}\},\quad i \in \{1,2,...,\dim(\mathbf{Zm})\} \tag{138}$$

$$\lambda m_L\big|_{+1} = \upsilon\{\mathrm{L}_{j+1}\big|_j, ... , \mathrm{L}_{j+k+1}\big|_{j+k}\} \tag{139}$$

É interessante notar que o uso de análogos observados conduz à dramática consequência de inexistência de garantia de variabilidade nula, mesmo na ausência de perturbações, ao contrário do enunciado na Equação (135). De fato, esta condição configura um caso particular da condição genérica descrita nas Equações (140-141), onde o limite superior de probabilidade apresentado corresponde à execução perfeita. A ausência de garantia de variabilidade nula, ainda que sob a restritiva condição de ausência de perturbações não medidas, se origina na diferença entre a previsão do estado subseqüente, $\mathbf{Zm}_{j+1}\big|_j$, e a apropriação das informações realizada ao cabo da implementação das soluções ótimas, $\mathbf{Zm}_{j+1}$. Esta discordância será projetada, pelo RTO, no espaço reduzido das variáveis atualizadas, onde ela será explicanda unicamente à luz do vetor de adaptação $\boldsymbol{\theta} = \boldsymbol{\Theta}(\mathbf{upd})$. Isto causará mudanças nas previsões do estado ótimo e do desempenho no estado consecutivo, do modo como exemplificado na Equação

(142). Note-se que no fluxo de informações considerado na Equação (142) está implícito que a solução proposta pelo RTO é implementada, ou seja, $\mathbf{Zm}_{j+1}(\mathbf{df})\big|_j \rightarrow \mathbf{Z}_{j+1}(\mathbf{df})$.

$$p\left((\boldsymbol{\lambda}\mathbf{m}_Z = \mathbf{0})\big|_{\mathbf{Z(d)=0}}\right) \leq 1 \tag{140}$$

$$p\left((\boldsymbol{\lambda}\mathbf{m}_Z\big|_{+1} = \mathbf{0})\big|_{\mathbf{Z(d)=0}}\right) \leq 1 \tag{141}$$

$$\left\{\mathbf{Zm}_{j+1}(\mathbf{ms})\big|_j \neq \mathbf{Zm}_{j+1}(\mathbf{ms})\right\}_{\mathbf{Z}_{j+1}(\mathbf{dum})=\mathbf{Z}_j(\mathbf{dum})} \Rightarrow \left\{\boldsymbol{\theta}_{j+1} \neq \boldsymbol{\theta}_j\right\} \Rightarrow$$
$$\left\{\mathbf{Zm}_{j+2}(\mathbf{df})\big|_{j+1} \neq \mathbf{Zm}_{j+1}(\mathbf{df})\big|_j\right\} \Rightarrow \left\{\mathbf{Zm}_{j+2}\big|_{j+1} \neq \mathbf{Zm}_{j+1}\big|_j\right\} \Rightarrow \left\{\mathbf{L}_{j+2}\big|_{j+1} \neq \mathbf{L}_{j+1}\big|_j\right\} \tag{142}$$

Na Seção 3.1.1.1 serão apresentados alguns exemplos que demonstram a existência das consequências previstas nas Equações (140-142). Por ora, seria relevante reunir elementos que expliquem os motivos pelos quais $\mathbf{Zm}_{j+1}(\mathbf{ms})\big|_j \neq \mathbf{Zm}_{j+1}(\mathbf{ms})$, que é o fator condicionante da variabilidade observada nas tomadas de decisão do RTO, assim como no desempenho a elas associado.

Para melhor responder esta pergunta é necessário ter em mente a real natureza das intervenções do sistema de RTO, que ficam explícitas na Figura 4, onde se evidencia os condicionantes associados à operação em malha fechada. Um primeiro motivo seria que o processo tenha sido submetido a novas condições, impostas após a tomada de decisão ótima, ou seja, $\mathbf{Z}_{j+1}(\mathbf{d}) \neq \mathbf{Z}_j(\mathbf{d})$. Este é um risco permanente, tornado mais acentuado se o RTO é estático, dado o intervalo de tempo decorrido até que a estacionariedade seja atingida. Este risco está associado a fatores além do controle da operação do RTO, embora sua frequência de ocorrência possa ser previamente investigada. Contudo, na vigência de condições de completude de informações e de seu perfeito processamento, esta condição não causaria a cadeia de eventos em (142), desde que preservada a constância das perturbações não medidas.

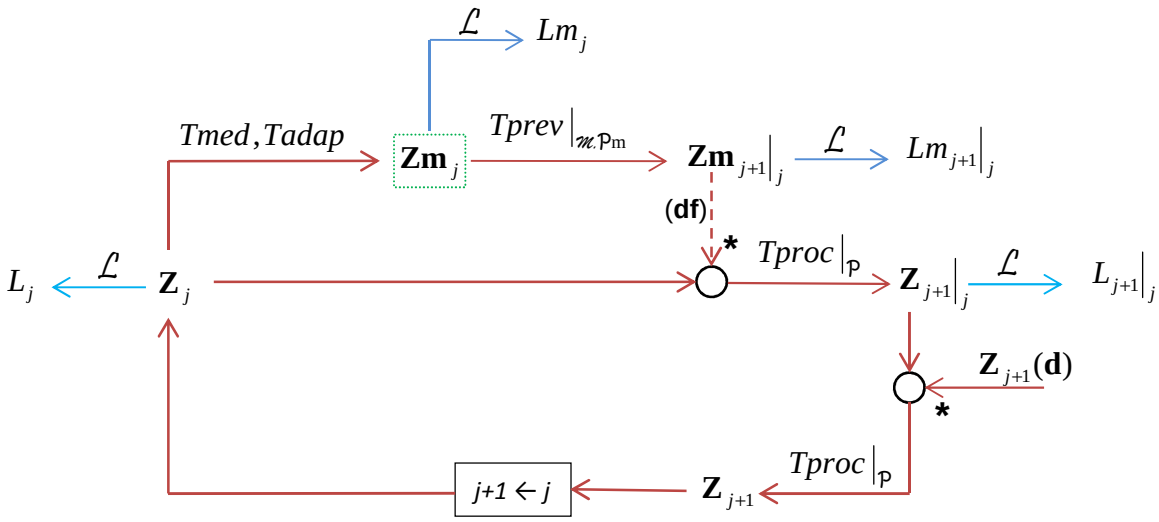


Figura 4 – Otimização em duas etapas. Comportamento em malha fechada. As linhas com asterisco indicam que os elementos indicados substituirão seus análogos no ponto de junção.

Um segundo motivo está relacionado ao processo de medição, em virtude do caráter probabilístico das observações diretas (conforme Equação (79)). Como este fato traduz uma realidade exterior ao processo, sua influência sobre a variabilidade poderia, em tese, ser anulada na medida em que o processo de adaptação impedisse a sua propagação para a camada de otimização de desempenho. Seria isto possível, mesmo em condições idealizadas?

Para que isto ocorra de forma, são necessárias duas condições: que todas as variáveis medidas cujo sinal seja passível de corrupção sejam reconciliadas, **rec**=**crp** (Equação (84)), e que o método de adaptação seja capaz de fielmente resgatar em **Zm** a informação associada à **Z**(**upd**). Supondo-se que tal método perfeito exista, no mundo real ocorre que todas as informações são corrompidas (**crp** = **ms).** Para permitir que o processo de adaptação supere este fato, é necessário que todas as informações medidas sejam reconciliadas, **rec**=**ms.** Isto implica que todas as informações medidas sejam apropriadas pelas funções de medição, $\mathbf{ms}^{-}$ = **ms,** e, por consequência, **dual** = ∅ (Equações 80-81), violando o previsto na Equação (110), que prevê que a função objetivo do problema de adaptação deva conter ao menos uma variável dual.

Em vista deste fato, o melhor que se pode fazer é realizar de forma criteriosa as escolhas possíveis a respeito da estrutura do RTO de modo a minimizar os efeitos da não idealidade do processo de adaptação. O RTO deve prover mecanismos que façam com que os desvios incorporados em **Zm** o sejam com probabilidade menor do que aqueles incorporados em **Za** pelo processo de medição. Deste modo, dada a impossibilidade de

anular integralmente os efeitos da corrupção dos sinais medidos, resta ao RTO minimizar a razão *r* expressa na Equação (143).

$$r = \frac{p\left(\left|\mathbf{Za}_j(\mathbf{ms}) - \mathbf{Z}_j(\mathbf{ms})\right| < \Delta\right)}{p\left(\left|\mathbf{Zm}_j(\mathbf{ms}) - \mathbf{Z}_j(\mathbf{ms})\right| < \Delta\right)} \tag{143}$$

O terceiro motivo está relacionado ao processamento incorreto da informação e diz respeito aos fatores que impedem a consecução do processo descrito na Equação (72). Como discutido na Seção 2.2.2, o rol de causas deste fenômeno está além do uso de relações estruturais inapropriadas ('model mismatch'), e diz respeito também à incompletude e inadequação do conjunto de informações disponível, bem como à sub-representação das variáveis pertinentes (cujo caso exemplar está discutido na Seção 2.5). É importante ressaltar que estes fatores, diferentemente da natureza estocástica das medições, além de adicionarem variabilidade inútil às intervenções do RTO, muito provavelmente condenarão os resultados a serem confinados em níveis sub-ótimos, o que é uma terminologia neutra que abarca, dentro de seu significado, a possibilidade de perspectivas menos brandas, como o fato de o RTO causar intervenções que tornem pior o desempenho do processo em relação ao seu estado imediatamente anterior.

Um fato particularmente difícil neste aspecto é que o RTO em duas etapas não dispõe de mecanismos para se proteger de tais riscos. Suas salvaguardas estão depositadas no procedimento de adaptação, que não é projetado para lidar com estes tipos de dificuldades pois está limitado a manipular os fatores aditivos $\boldsymbol{\Theta}$. Um modo fácil de enxergar esta situação é perceber que, no impedimento do fluxo de informações previsto na Equação (72), a variabilidade prevista na Equação (142) se manifestará, ainda que na presença de medições determinísticas e acuradas.

Biegler, Grossman e Westerberg [58] lidaram com um problema aparentemente distinto do RTO, mas cuja análise é diretamente aplicada ao caso presente. À época, por questões de carga computacional, ganhou popularidade, em problemas de simulação de equilíbrio multicomponente, o uso do método “inside-out”, que prevê a resolução do problema em duas camadas hierarquizadas, associadas a modelos de graus distintos de complexidade. A idéia subjacente consiste em resolver os cálculos iterativos distribuindo a sua maior parte para um modelo simplificado, contido no laço interno de iterações. O modelo rigoroso forma um laço externo que é alimentado com a solução proposta pelo

modelo simplificado. As respostas do modelo rigoroso são utilizadas para atualizar os parâmetros do modelo simplificado. Novo ciclo de iterações do laço interno é iniciado a partir desta atualização e o processo continua ciclicamente até que a convergência seja atingida. A analogia funcional com o problema de otimização apresentado na Figura 4 é direta, devendo-se levar em conta que a planta real faz o papel do modelo rigoroso.

A condição necessária para o reconhecimento do ponto ótimo pelo modelo simplificado é que os gradientes das relações de igualdade e desigualdade do modelo e do processo sejam idênticos no ponto ótimo para todas as variáveis. A condição suficiente é que tal relação seja observada para todos os pontos [40], ou seja, $\nabla \mathbf{g} = \nabla \mathbf{gm}$, $\nabla f = \nabla fm$, o que implicitamente considera que todas as variáveis do processo real estão contidas no modelo (**Zm**=**Z**). A (extremamente provável) não observância desta condição implica que o método de otimização em duas camadas necessariamente levará mais de um ciclo até a estabilização e que a condição de otimalidade não será atingida.

3.1.1.1. Exemplos de causas da variabilidade

Esta Seção tem por finalidade apresentar casos que exemplifiquem didaticamente os conceitos de variabilidade das informações produzidas pelo RTO, conforme descritos em 3.1.1, realçando não só os esperados efeitos de variabilidade devido à corrupção da informação, como também devido ao processamento incorreto da informação, como descrito na Seção 2.2.2.

Os dois primeiros casos aqui descritos evidenciam conseqüências do processamento incorreto da informação (vide pg. 37) . O caso 1 demonstra a ocorrência de processamento incorreto dos tipos 3 e 4, enquanto que o caso 2 apresenta um exemplo de processamento incorreto do tipo 1. O terceiro caso mostra a variabilidade associada à presença de corrupção da informação, na ausência de outras condições que afastem o sistema das condições necessárias à perfeita execução do RTO. Um resumo das características dos três casos pode ser visto na Tabela 2.

Tabela 2 – Resumo das características de cada caso

Caso 1	Caso 2	Caso 3
$\boldsymbol{fm} \neq \boldsymbol{f}$	$\boldsymbol{fm} = \boldsymbol{f}$	$\boldsymbol{fm} = \boldsymbol{f}$
	$\mathbf{Zm}_0 \neq \mathbf{Z}_0$	$\mathbf{Zm}_0 = \mathbf{Z}_0$
crp = [ ]	**crp** = [ ]	**crp** = **ms**

O processo que constituirá o tema a partir do qual as variações serão contruídas será tomado de empréstimo daquele originalmente apresentado por Biegler *et alii* [58] e também usado por and Zhang e Forbes [4]. A função objetivo é apresentada na Equação (144), enquanto que a relação funcional de igualdade para o processo real é dada pela Equação (145). O conjunto de equações através do qual o RTO irá operar é resumido na Equação (146), consistindo no modelo do processo, $\boldsymbol{fm}$, nas funções de atribuição, $\boldsymbol{f}_{\boldsymbol{atr}}$, e nas relações derivadas do processo de medição, $\boldsymbol{f}_{\boldsymbol{med}}$, sendo seus componentes detalhados individualmente na descrição específica de cada caso.

$$L = (y - 1/2)^2 + (x-1)^2 \tag{144}$$

$$\boldsymbol{f}\text{:}\quad (x-c_3)^3 + (x-c_2)^2 + c_1 - y = 0 \tag{145}$$

$$f_{sis}\begin{cases} f \;\; : \boldsymbol{fm} \\ f_{atr} : \mathbf{Zm}_j(\mathbf{atr}) = \mathbf{Zm}_0(\mathbf{atr}) + \boldsymbol{\Theta}_j(\mathbf{atr}) \\ f_{med} : \mathbf{Zm}(\mathbf{ms}^-) = \mathbf{Zm}(\mathbf{ms}^-) + \boldsymbol{\Theta}_j(\mathbf{ms}^-) \end{cases} \tag{146}$$

De acordo com a representação proposta neste texto, as variáveis que descrevem a planta em (145) e suas respectivas condições de partida são apresentadas, respectivamente, nas Equações (147,148):

$$\mathbf{\{Z\}} = \{x, c_3, c_2, c_1, y\}\ ,\mathbf{df}_z = 1 \tag{147}$$

$$\mathbf{Z}_0 = [1, 1, 1, 1, 1]^T \tag{148}$$

O cenário de operação do RTO compõe-se de 10 ciclos consecutivos de intervenção do otimizador em malha fechada, sendo caracterizado pela manutenção das condições da planta (Equações 149,150), exceção feita de possíveis manipulações pelo RTO e de seus impactos na variável consequente, *y*.

$$\mathbf{Zcen} = [\mathbf{Z}_0 \ldots \mathbf{Z}_{Ncen-1}],\ \mathbf{Z}_j = \mathbf{Z}_0\ \forall\, j,\ Ncen{=}10 \tag{149}$$

$$\mathbf{fix} = [1, 2, 3, 4, 5]^T\ ;\ \mathbf{var}\ =\ \phi \tag{150}$$

Caso 1:

Neste caso, é suposto o processamento incorreto do tipo 3 (Equação 74) uma vez que o modelo descreve a planta por meio da relação estruturalmente imperfeita mostrada na Equação (151), cujas variáveis descritivas contidas em **Zm** e o estado de conhecimento inicial do processo, $\mathbf{Zm_0}$, são apresentados, respectivamente, nas Equações (152) e (153).

$$\boldsymbol{fm}\text{: } x + \beta - y = 0 \tag{151}$$

$$\mathbf{\{Zm\}} = \{x, \beta, y\}\ ,\mathbf{df} = 1 \tag{152}$$

$$\mathbf{Zm_0} = [1, 0, 1]^T \tag{153}$$

O conjunto de variáveis necessárias escolhido para a representação do processo é dado por {**in**} = {x, β}, de modo que **in** = $[1, 2]^T$, **out** = [3].

As variáveis x e y são obtidas por observação direta sem corrupção da informação (Equação 154), sendo que x pertence a $\mathbf{ms}^-$, fazendo parte das funções de medição, $\boldsymbol{f}_{\text{med}}$, enquanto que y é variável dual (Equação 155). A variável β é indiretamente observada (Equação 156), sendo manipulada pelo processo de adaptação.

$$\mathbf{ms} = [1,3];\ \mathbf{crp} = \phi \tag{154}$$

$$\mathbf{ms}^- = [1],\ \mathbf{dual} = [3] \tag{155}$$

upd = [2]; **est** = [2]; **rec** = ϕ (156)

A seqüência de ações do RTO pode ser resumida, para este caso, através dos passos abaixo descritos e representados nas Equações (157-160,166):

1) Dadas as variáveis necessárias previstas em cada ciclo do cenário de operação, a planta produz as variáveis conseqüentes mediante atendimento do modelo do processo:

$$\mathbf{Z}_j(\mathbf{out}) : \left\{ \mathbf{y} \,\middle|\, \boldsymbol{f}\left(\mathbf{Z}_j \xleftarrow{(\mathbf{out})} \mathbf{y}\right) = \mathbf{0} \right\} \tag{157}$$

2) Em virtude da apropriação perfeita da informação, as variáveis obtidas por observação direta (**ms**) espelham seus análogos em **Z** (**ms**$_z$):

$$\mathbf{Zm}_j(\mathbf{ms}) = \mathbf{Z}_j(\mathbf{ms}_z) \tag{158}$$

3) O parâmetro β é calculado de modo a explicar **Zm**$_j$(**ms**) à luz do modelo ***fm***:

$$\mathbf{Zm}_j(\mathbf{est}) : \left\{ \boldsymbol{\beta} \,\middle|\, \boldsymbol{f}(\mathbf{Z}_j) = \boldsymbol{fm}(\mathbf{Zm}_j \xleftarrow{(\mathbf{est})} \boldsymbol{\beta}) = \mathbf{0} \right\} \tag{159}$$

4) O valor ótimo da variável de decisão previsto para vigorar a partir do próximo ciclo é calculado de modo a atender a condição de optimalidade

$$\mathbf{Zm}_{j+1}(\mathbf{df})\big|_j : \left\{ \mathbf{x} \,\middle|\, \nabla_{\mathbf{u}} L(\mathbf{Zm})\big|_{\boldsymbol{fm}(\mathbf{Zm})=\mathbf{0}} = \mathbf{0},\ \det\left(H_{\mathbf{u}}\left(L(\mathbf{Zm})\big|_{\boldsymbol{fm}(\mathbf{Zm})=\mathbf{0}}\right)\right) > \mathbf{0},\ \mathbf{Zm}_{j+1} \xleftarrow{(\mathbf{df})} \mathbf{x} \right\} \tag{160}$$

No presente caso, em que dim(**u**) = dim(**df**) = 1, esta condição assume a forma mais simples expressa nas Equações (161) a (165):

$$\mathbf{Zm}_{j+1}(\mathbf{df})\big|_j : \left\{ \mathbf{x} \,\middle|\, \frac{\partial Lm}{\partial x} = 0,\ \frac{\partial^2 Lm}{\partial x^2} > 0,\ \mathbf{Zm}_{j+1} \xleftarrow{(\mathbf{df})} \mathbf{x} \right\} \tag{161}$$

$$Lm = L(\mathbf{Zm})\Big|_{fm(\mathbf{Zm})=0} \tag{162}$$

$$Lm = \left(\underbrace{x + \beta}_{y} - 1/2 \right)^2 + (x-1)^2 \tag{163}$$

$$\frac{\partial\, Lm}{\partial\, x} = 4x + 2\beta - 3 \tag{164}$$

$$\frac{\partial^2\, Lm}{\partial\, x^2} = 4 \tag{165}$$

5) A solução do RTO é implementada para o próximo ciclo:

$$\mathbf{Z}_{j+1}(\mathbf{df}) \leftarrow \mathbf{Zm}_{j+1}(\mathbf{df})\Big|_j \tag{166}$$

Para o processo *verdadeiro*, a expressão analítica da função objetivo em função das variáveis necessárias é expressa pela relação (167), obtida por meio da substituição do valor de y em (145) na Equação (144). As expressões das derivadas de primeira e segunda ordens da função objetivo são mostradas nas Equações (168,169).

$$L = \left(\underbrace{x^3 - 3c_3x^2 + 3c_3^2x - c_3^3 + x^2 - 2c_2x + c_2^2 + c_1}_{y} - 1/2 \right)^2 + (x-1)^2 \tag{167}$$

$$\frac{\partial\, L}{\partial\, x} = 2a\,b + 2x - 2 \tag{168}$$

$$\frac{\partial^2\, L}{\partial\, x^2} = 2a(6x - 6c_3 + 2) + 2b(3x^2 - 6c_3x + 3c_3^2 + 2x - 2c_2) + 2 \tag{169}$$

onde

$a = x^3 - 3c_3x^2 + 3c_3^2x - c_3^3 + x^2 - 2c_2x + c_2^2 + c_1 - 1/2$

$b = 3x^2 - 6c_3x + 3c_3^2 + 2x - 2c_2$

No presente caso, os resultados das sucessivas implementações do RTO segundo os cinco passos anteriormente descritos podem ser vistos na Tabela 3, onde os valores das derivadas foram calculados analiticamente por meio das Equações (164,168,169). Note-se que o ponto de partida, referenciado pelas condições no ciclo 1, já é o ponto de melhor desempenho possível do sistema, assinalado pelos valores das primeira e segunda derivadas de L em relação a x. O resultado efetivo da ação do RTO é o de adicionar variabilidade ao processo e conduzí-lo a um novo estado sub-ótimo, nunca retornando às condições de optimalidade presentes no ciclo 1.

Tabela 3 - Comportamento do modelo e da planta ao longo de sucessivas intervenções do RTO para o caso 1

Ciclo	x_j	β_j	y_j	$\partial Lm/dx$	$\partial L/dx$	$\partial^2 L/dx^2$
1	1	0	1	1	**0**	**4**
2	0.75	0.30	1.05	0.59	-0.84	2.74
3	0.60	0.49	1.10	0.39	-1.18	1.72
4	0.50	0.62	1.12	0.25	-1.31	0.88
5	0.44	0.70	1.14	0.15	-1.35	0.33
6	0.40	0.74	1.14	0.09	-1.35	-0.03
7	0.38	0.77	1.15	0.05	-1.35	-0.21
8	0.37	0.78	1.15	0.03	-1.35	-0.29
9	0.36	0.79	1.15	0.01	-1.35	-0.38
10	0.36	0.79	1.15	0.01	-1.34	-0.38
11	0.35	0.79	1.15	0.00	-1.34	-0.46

Neste exemplo há constância do cenário de operação (Equações 148,149) e apropriação perfeita da informação obtida por observação direta (Equação 154). Contudo, conforme descrito na Seção 3.1.1, mesmo na ausência de variabilidade externa, o RTO induz, de forma autônoma, a variabilidade das condições operacionais e da função objetivo, conforme visto nas Figuras 5 e 6.

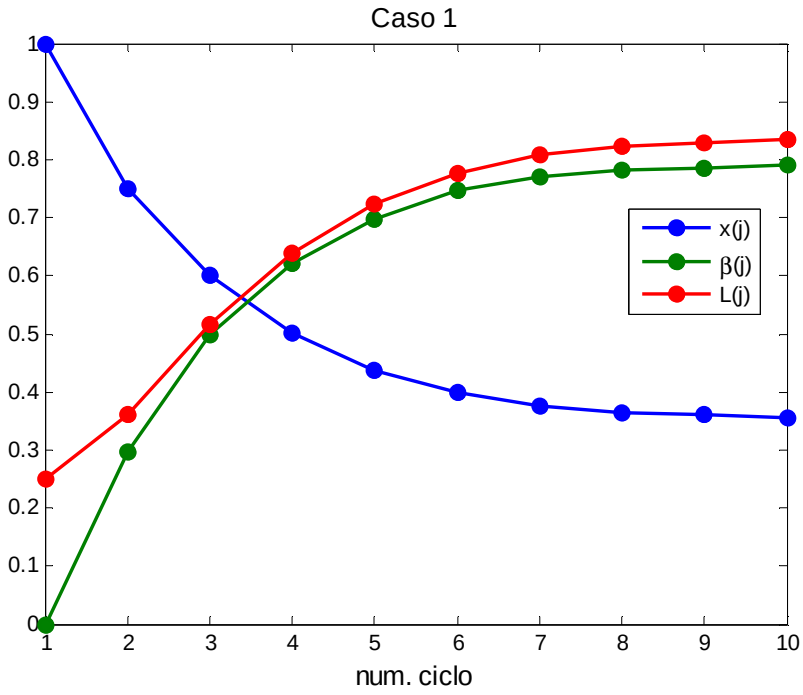


Figura 5 - Comportamento da variável de decisão, x, da variável adaptada, β, e da função objetivo, L, ao longo do cenário de operação do RTO para o caso 1.

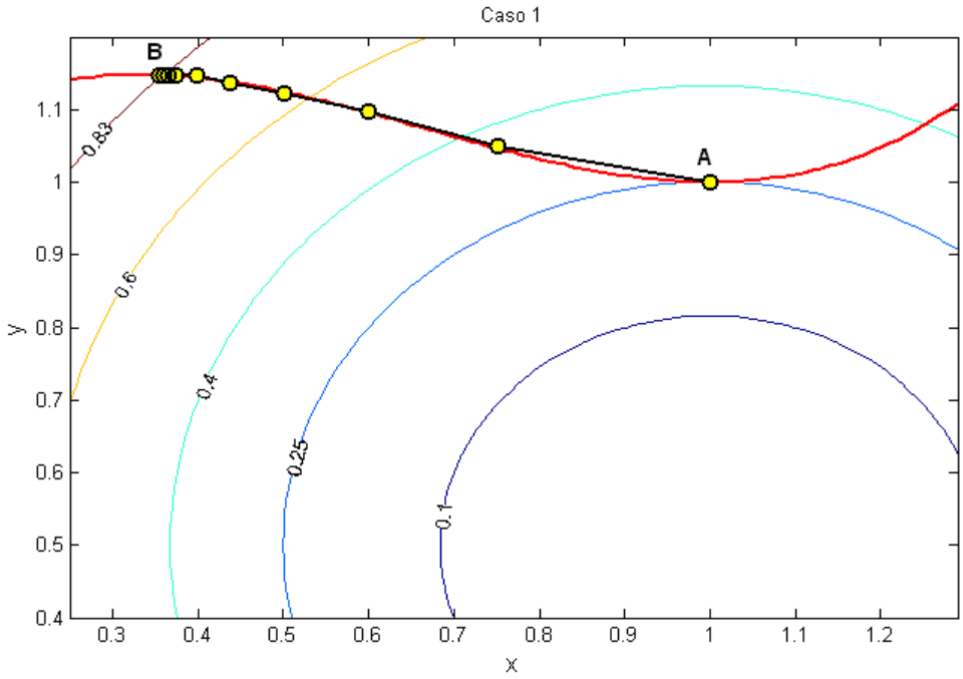


Figura 6 - Trajetória do comportamento do processo ao longo do trajeto (A→B) em virtude da sucessiva implementação em malha fechada das soluções do RTO. Curvas de nível: função objetivo; Curva vermelha: modelo real do processo. Círculos amarelos: (x_j,y_j)

Algumas características da variabilidade induzida pelo RTO em caso de processamento incorreto da informação merecem ser reunidas nas seguintes propriedades:

P1) a ausência de variabilidade externa ao processo e a ausência de corrupção das informações não garantem ausência de variabilidade do processo sob intervenção de um sistema de RTO, mesmo se as medições forem perfeitas:

$$Se \quad \exists \mathbf{u} : \left\{ \nabla_{\mathbf{u}} \boldsymbol{L}(\mathbf{Zm}) \big|_{\boldsymbol{fm}(\mathbf{Z}m)=\mathbf{0}} = \mathbf{0};\ \nabla_{\mathbf{u}} \boldsymbol{L}(\mathbf{Z}) \big|_{\boldsymbol{f}(\mathbf{Z})=\mathbf{0}} \neq \mathbf{0} \right\}$$

Então $\{\mathbf{Zm}_j(\mathbf{ms}) = \mathbf{Z}_j(\mathbf{ms}_z), \mathbf{Z}(\mathbf{d}) = 0\} \not\Rightarrow \boldsymbol{\lambda}\mathbf{m}_Z = 0$

P2) não há garantia de optimalidade. As variáveis realizam um trajeto determinístico cuja estabilização não está comprometida com o ótimo da função objetivo principal. De fato, neste caso, o sistema estava em seu ponto ótimo e foi dele deslocado por conta exclusiva da ação do RTO;

estabilização: $\left\{ \mathbf{Z}_{j+1}(\mathbf{ms}_z) = \mathbf{Z}_j(\mathbf{ms}_z) \right\} \not\Rightarrow \nabla_{\mathbf{u}} \boldsymbol{L}(\boldsymbol{Z}) \big|_{\boldsymbol{f}(\mathbf{Z})=\mathbf{0}} = 0$

P3) o RTO cessa a indução de variabilidade em um número de ciclos que é função da configuração do problema, a saber: {***f***, ***f*****m**, $\mathbf{Z_0}$, $\mathbf{Zm_0}$}. Em termos reais, em função da frequência de intervenção do RTO e da dinâmica do processo, estes ciclos podem representar um longo tempo. Para o caso genérico não linear, não há como formular *a priori* uma regra de previsibilidade do formato nem da duração da trajetória até a estabilidade.

Mesmo para um problema simples, como o caso atual, a multiplicidade de trajetórias passíveis de serem percorridas pode ser considerável. A alteração de qualquer elemento pertinente à configuração do problema pode acarretar mudanças consideráveis no formato das trajetórias.

Como exemplo, pode-se observar os resultados contidos nas Figuras 7 e 8, que mostram a influência da condição do processo anterior às intervenções do RTO, representada por x_0. Nestas figuras, o impacto de x_0 sobre a evolução do processo foi

aferido segundo dois pontos de vista: 1) a duração da trajetória do processo até a estabilidade, medida com o auxílio do parâmetro, Δ_{est}, que contempla a estabilização simultânea da variável de decisão, da variável estimada e da função objetivo, sendo definido na Equação (170); 2) a variabilidade induzida pelo RTO, medida por meio do parâmetro λ_e, definido na Equação (171), que registra o valor normalizado do desvio padrão dos estados de cada variável ao longo da trajetória até a estabilidadede.

$$\Delta_{est} = \left\{ j \mid 100\frac{(a_j - a_{j-1})}{a_{j-1}} < tol, \quad a \in \{x, \beta, L\} \right\} \tag{170}$$

$$\lambda_e = \sigma(a_{j=2..ne}) / a_{ne}, \quad a \in \{x, \beta, L\} \tag{171}$$

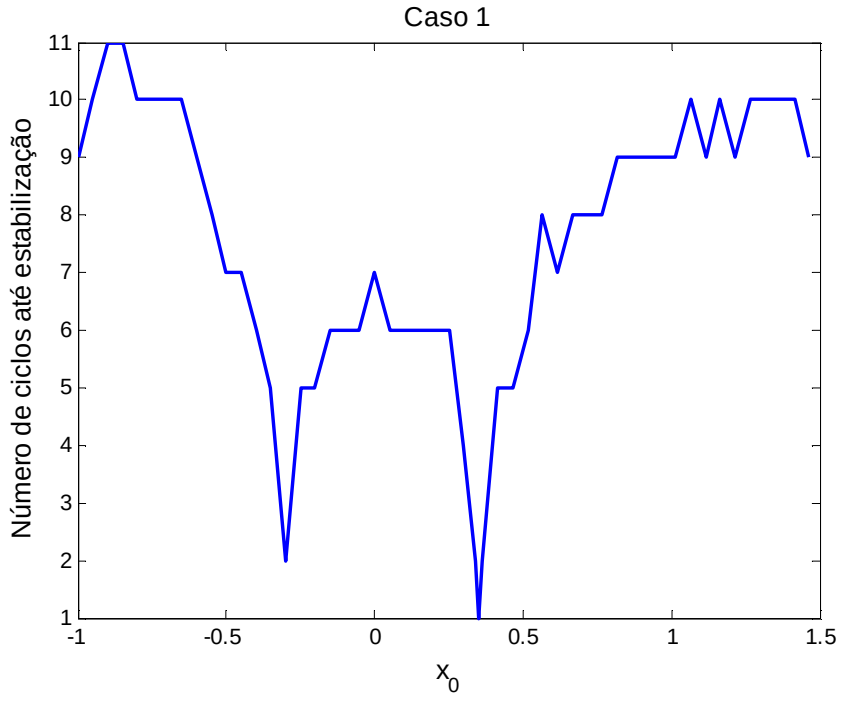


Figura 7 – Influência do valor inicial de x antes do início das intervenções do RTO sobre a duração do período de estabilização dos valores de processo. tol = 2% (vide texto)

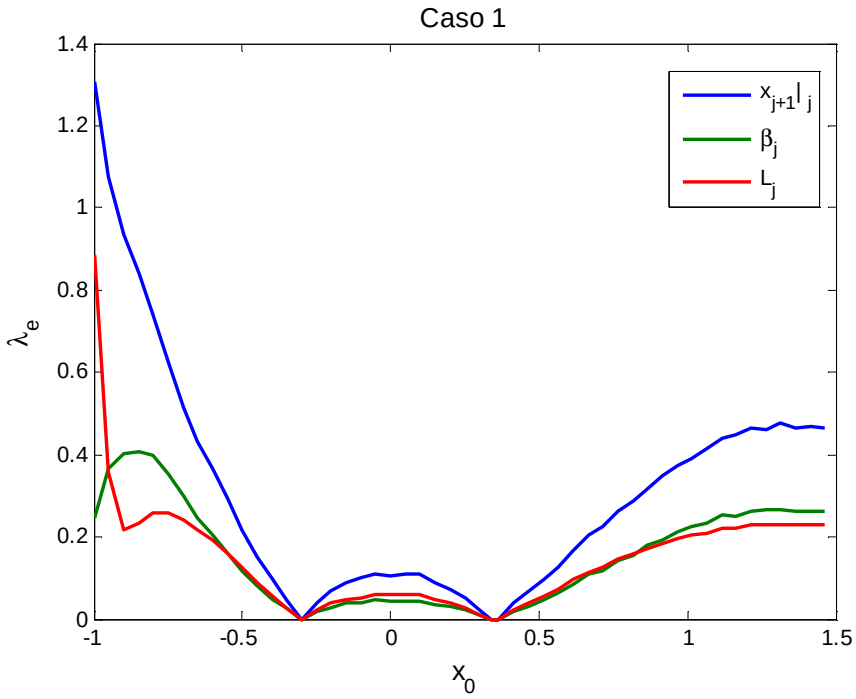


Figura 8 – Medida de variabilidade do caso 1 em função do valor inicial de x para três variáveis representativas.

Uma vez que, no presente caso, as funções e as derivadas podem ser descritas analiticamente, além do que as informações obtidas por observação direta espelham perfeitamente o comportamento real, pode-se verificar explicitamente o modo como os fatos observados nas Figuras 5 a 7 são produzidos. Assim sendo, pode-se formular o resultado do problema de otimização expresso na implementação prevista para a variável de decisão a ser implementada no próximo ciclo, x_{j+1}, em função das condições presentes no ciclo j. Este formulação é consequência do atendimento das condições de optimalidade (Equações 161, 164,165) e é expressa na Equação (172) :

$$x_{j+1} = \frac{3 - 2\beta_j}{4} \tag{172}$$

Da mesma forma, pode-se implementar o procedimento de estimação previsto na Equação (159) para representar a formulação de β explicitando a dependência do valor da variável de decisão implementado no próximo ciclo com seu valor no ciclo atual (Equações 173,174). Assim, por meio da combinação das Equações (172) e (174), chega-se à relação de dependência entre dois valores consecutivos da variável de decisão, apresentado na Equação (175).

$$\beta_j = x_j^3 + (1-3c_3)x_j^2 + (3c_3^2 - 2c_2 - 1)x_j + c_2^2 - c_3^3 + c_1 \quad (173)$$

$$\beta_j\big|_{\mathbf{Z}=\mathbf{Z}_0} = x_j^3 - 2x_j^2 + 1 \quad (174)$$

$$x_{j+1} = -\frac{x_j^3}{2} + x_j^2 + \frac{1}{4} \quad (175)$$

O valor da variável de decisão para o qual o RTO converge no presente caso corresponde à raiz da Equação (175) sob a condição $x_{j+1} = x_j$. Tal condição corresponde ao valor apresentado na Equação (176), que é igual ao valor proposto empiricamente pelo RTO, como verificado na Tabela 3 e nas Figuras 5 e 6.

$$x_{j+1} = x_j \;\Rightarrow\; x_j \approx 0{,}3522 \quad (176)$$

Caso 2:

Diferentemente do caso 1, no exemplo atual o modelo de representação da planta é estruturalmente idêntico à realidade, de modo que ***fm*** = ***f***, como apresentado na Equação (177),

$$\boldsymbol{fm}: \;\; (x-c_3)^3 + (x-c_2)^2 + c_1 - y = 0 \quad (177)$$

O conjunto de variáveis representado no modelo também é idêntico ao real, $\{\mathbf{Zm}\} = \{\mathbf{Z}\}$. O conjunto de variáveis necessárias escolhido para a representação do processo é dado por $\{\mathbf{in}\} = \{x, c_3, c_2, c_1\}$, de modo que $\mathbf{in} = [1, 2, 3, 4]^T$, $\mathbf{out} = [5]$.

As variáveis x e y são obtidas por observação direta sem corrupção da informação (178), sendo que x faz parte das funções de medição, $\boldsymbol{f}_{med}$, enquanto que y é variável dual (179). A variável c_3 é indiretamente observada (180), sendo obtida por meio do processo de adaptação.

As variáveis c_3, c_2 e c_1 fazem parte das funções de atribuição, $\boldsymbol{f}_{atr}$. Contudo, apenas c_3 e c_2 fazem parte do conjunto de informações *a priori* verdadeiras (181), fato que também pode ser depreendido da análise comparativa das Equações (148) e (182), que evidenciam que $\mathbf{Z_0} \neq \mathbf{Zm}_0$. Deste modo, configura-se um caso de processamento incorreto da informação do tipo 1 (vide Seção 2.2.2). O resumo da classificação das variáveis em Zm em termos de sua função no RTO é mostrado na Tabela 4.

$$\mathbf{ms} = [1,5];\ \ \mathbf{crp} = \phi \tag{178}$$

$$\mathbf{dual} = [5],\ \ \mathbf{ms}^- = [1] \tag{179}$$

$$\mathbf{upd} = [2];\ \ \mathbf{est} = [2];\ \mathbf{rec} = \phi \tag{180}$$

$$\mathbf{atr} = [2, 3, 4];\ \mathbf{apv} = [2, 3];\ \mathbf{apf} = [4];\ \mathbf{ocu} = \phi \tag{181}$$

$$\mathbf{Zm}_0 = [1, 1, 1, 0.9, 1]^T \tag{182}$$

Tabela 4 – Classificação das variáveis **Zm** para o RTO implementado no caso 2.

	x	c_3	c_2	c_1	y
in	●	●	●	●	
out					●
ms	●				●
ms$^-$	●				
dual					●
upd		●			
atr		●	●	●	
apv		●	●		
apf				●	
ocu					

Na Figura 9 é apresentado o comportamento do processo em virtude das sucessivas atuações do RTO em malha fechada ao longo do cenário de dez ciclos. Este exemplo, com a condição de partida $\mathbf{Zm}_0$ dada pela Equação (182) será chamado de caso 2a.

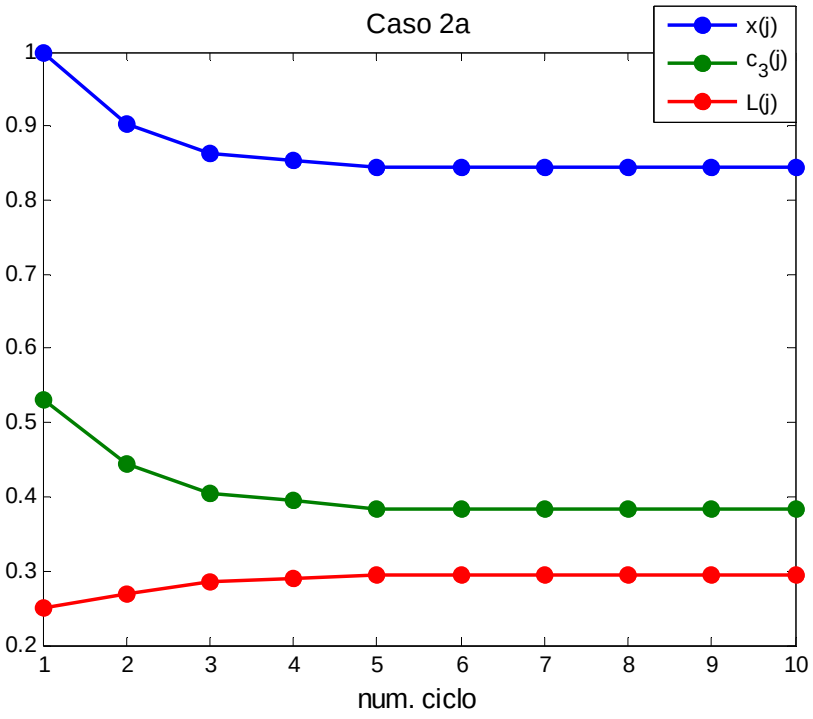


Figura 9 - Comportamento da variável de decisão, x, da variável adptada, c_3, e da função objetivo, L, ao longo do cenário de operação do RTO para o caso 2a.

Deve-se notar que, apesar de neste caso não haver diferença estrutural entre os modelos, ainda assim as mesmas características do caso 1 se apresentam, quais sejam: o dispêndio de mais de um ciclo para o RTO estabilizar suas decisões e o caráter sub-ótimo da situação de estabilidade. Estendendo um pouco mais o campo de observações para trazer à luz outra ocorrência de grande valor didático, pode-se testar uma variação deste caso de processamento incorreto da informação do tipo 1, alterando-se apenas o valor da condição *a priori* falsa associada a $\mathbf{Zm_0}$ conforme a Equação (183). Os resultados desta modificação podem ser observados na Figura 10. Esta modificação dará origem ao caso denominado 2b.

$$\mathbf{Z}_0 = [1, 1, 1, 1.1, 1]^T \qquad (183)$$

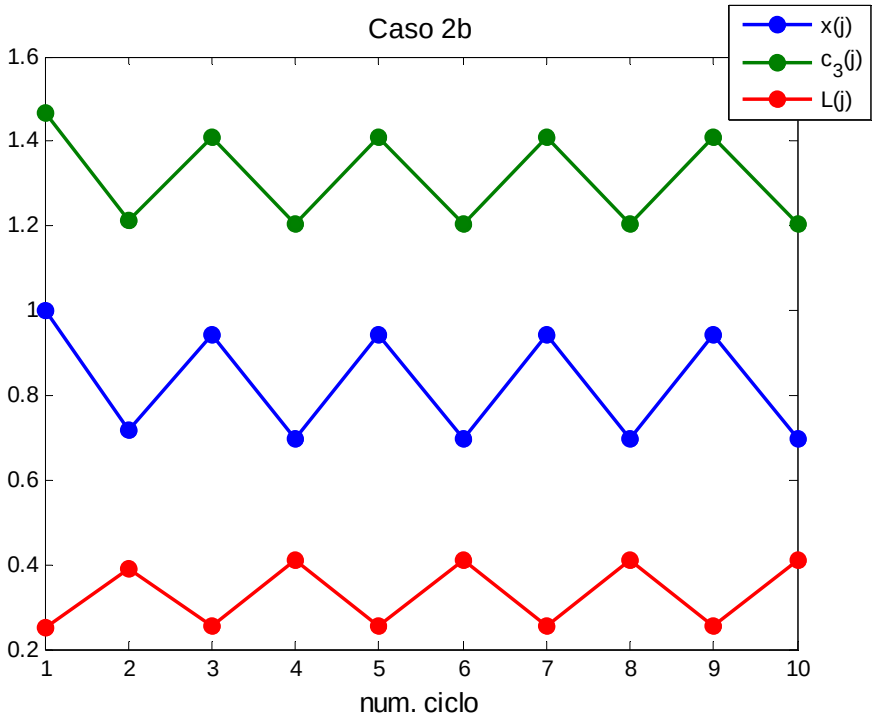


Figura 10 - Comportamento da variável de decisão, x, da variável adaptada, c_3, e da função objetivo, L, ao longo do cenário de operação do RTO para o caso 2b.

A comparação das Figuras 9 e 10 mostra quão importantes podem ser mudanças nas configurações do problema que, ao olhar não instrumentalizado, pareceriam de pequena relevância quantitativa. A condição $\mathbf{Z_0}$ assumida no caso 2b leva o RTO a adicionar variabilidade permanente ao sistema, fazendo-o oscilar continuamente entre dois estados.

A diferença de comportamento entre os casos 2a e 2b pode melhor ser acompanhada através da análise comparativa das Figuras 11 e 12. A construção do percurso do processo entre dois ciclos sucessivos pode ser visto da seguinte forma: a condição do processo no ciclo j é representada pelo ponto amarelo na interseção da linha vermelha que representa ***f***, de acordo com a Equação (145), com o modelo de representação para o ciclo j, dado pela linha azul tracejada 'y_j', que expressa ***fm*** atualizado de modo análogo à representação (159). A otimização de L produz ($x_{j+1}|_j$, $y_{j+1}|_j$), representado por um ponto verde ao longo da linha tracejada 'y_j'. O estado do processo no ciclo seguinte será indicado pelo ponto amarelo que assinala a projeção de ($x_{j+1}|_j$, $y_{j+1}|_j$) sobre ***f***, ao longo de $x_{j+1}|_j$.

Pode-se notar que, enquanto o ordenamento relativo dos modelos de representação (curvas tracejadas) para o caso 2a induz que estas triangulações produzam

estados sucessivos na mesma direção e sentido ao longo do eixo da variável de decisão, para o caso b este ordenamento induz o aprisionamento perpétuo do processo em uma região que, assintoticamente, será caracterizada por dois estados alternados.

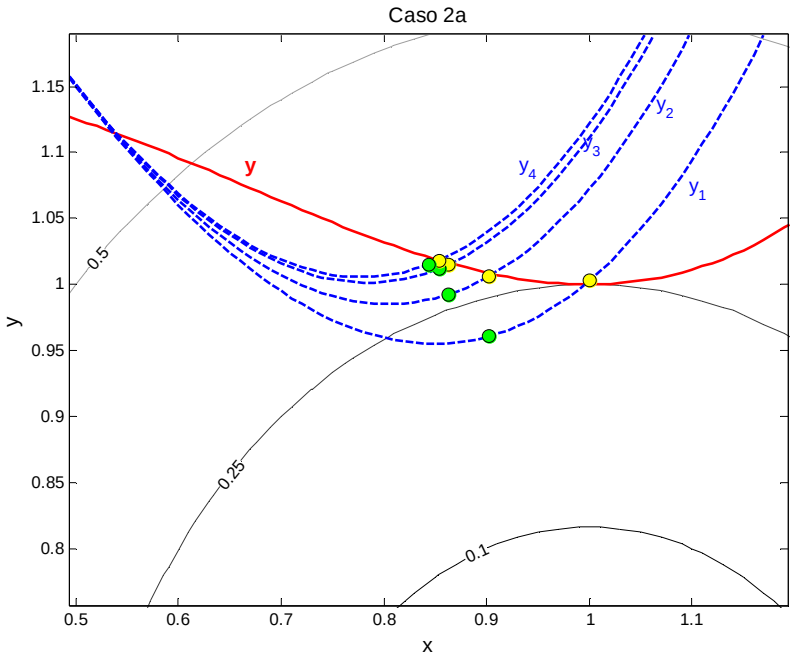


Figura 11 - Trajetória do comportamento do processo em virtude da implementação das soluções do RTO ao longo de quatro ciclos consecutivos. Curvas de nível: função objetivo; Curva vermelha: comportamento real do processo. Curvas azuis: y_j – modelo do processo no ciclo j em função de x. Círculos amarelos: (x_j,y_j). Círculos verdes: $(x_{j+1|j}, y_{j+1|j})$.

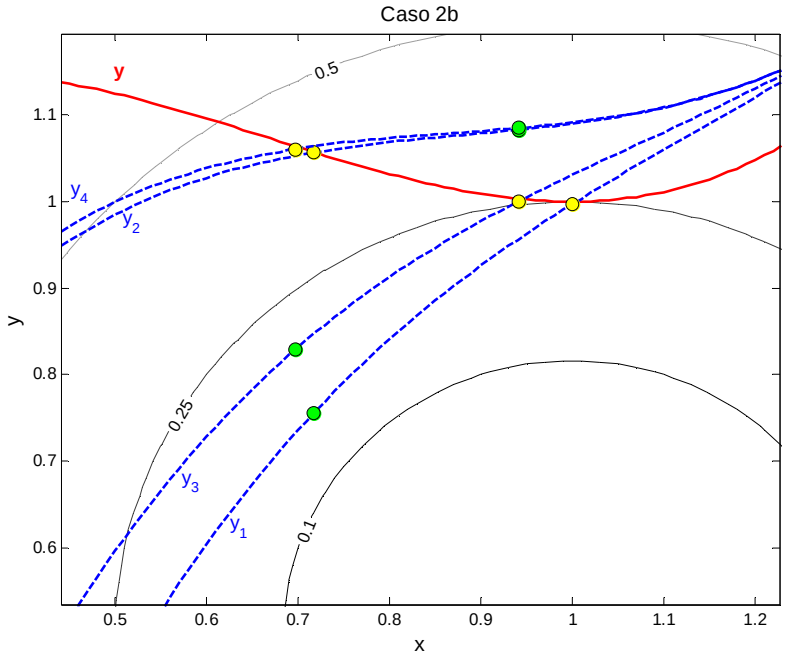


Figura 12 - Trajetória do comportamento do processo em virtude da implementação das soluções do RTO ao longo de quatro ciclos consecutivos. Curvas de nível: função objetivo; Curva vermelha: comportamento real do processo; Curvas azuis: y_j – modelo do processo no ciclo j em função de x. Círculos amarelos: (x_j,y_j). Círculos verdes: $(x_{j+1|j}, y_{j+1|j})$.

Caso 3:

Supõe-se que as variáveis x e y sejam obtidas por observação direta e que ambas estejam sujeitas à corrupção (Equação 184). Esta corrupção é fruto da contaminação aditiva da informação verdadeira com um sinal de natureza aleatória, como previsto na Equação (78), sendo que $\boldsymbol{\varepsilon}$ possui distribuição normal com média nula e desvio-padrão respectivamente igual a 3% do valor de cada variável em $\mathbf{Z_0}$.

O conjunto de variáveis necessárias escolhido para a representação do processo é dado por {**in**} = {x, c_3, c_2, c_1}, de modo que **in** = $[1, 2, 3, 4]^T$, **out** = [5]. A variável dual será y (Equação 185) e o RTO manipulará c_3 como forma de adaptação do modelo (186). Todas as variáveis de atribuição estão associadas a valores a priori verdadeiros (187) uma vez que $\mathbf{Zm_0} = \mathbf{Z_0}$ (Equação 188).

ms = [1,5]; **crp** = [1, 5] (184)

dual = [5], $\mathbf{ms}^{-}$ = [1] (185)

upd = [2]; **est** = [2]; **rec** = ϕ (186)

atr = [2, 3, 4]; **apv** = [2, 3, 4]; **apf** = ϕ; **ocu** = ϕ (187)

$\mathbf{Zm}_0 = \mathbf{Z_0} = [1, 1, 1, 1, 1]^T$ (188)

No presente caso, em que o processamento da informação é feito de modo perfeito, não é prevista, *a priori*, a indução da variabilidade determinística de baixa frequência ao longo dos ciclos, como verificado nos casos 1 e 2. Contudo, cada valor produzido pelo RTO estará associado a uma variável estocástica, como exemplificado pelos valores da variável de decisão, $x(j+1)|_j$, da variável estimada, $c_3(j+1)|_j$, e da função objetivo, L(j), como visto nas Figuras 13 e 14. Pode-se observar o afastamento das variáveis de decisão em relação ao seu valor ótimo (x_{otm} = 1), assim como o grau de sub-optimalidade, uma vez que o valor mínimo de L corresponde a 0,25.

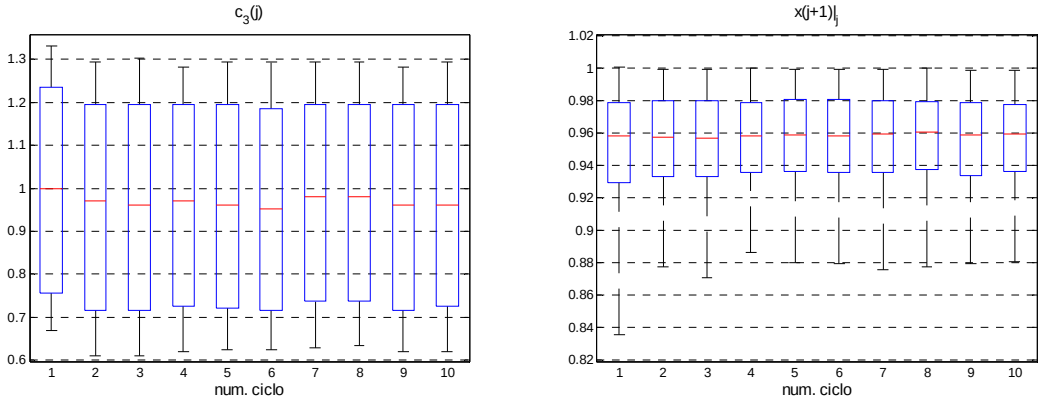


Figura 13 – Distribuição dos valores estimados (esquerda) e dos valores da variável de decisão (direita) do RTO para as condições do caso 3.

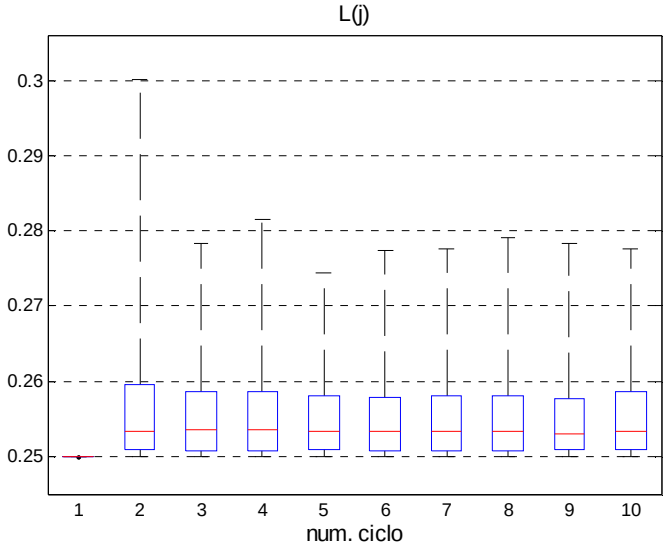


Figura 14 – Distribuição dos valores da função objetivo em cada ciclo antes da execução do otimizador para as condições do caso 3.

3.1.1.2. Questões relativas à estabilidade

Apesar de muito pouco explorado na literatura de otimização em tempo real, o tema da convergência das soluções propostas pelo RTO é de vital importância para o seu desempenho. Isto se dá na medida em que sistemas baseados nos procedimentos apresentados na Figura 4, fundamentados na realimentação das informações geradas, não possuem garantia prévia de que a série de soluções propostas S_u: (u_0, u_1, ..., u_n) convirja, ou seja, de que S_u seja uma série de Cauchy [59]. Isto corresponde ao fato de a relação (189) ser verdadeira uma vez definida a distância *d* no espaço métrico à qual pertence S_u.

$$\exists N : d(x_n - x_m) < \varepsilon, \quad \forall n,m > N \tag{189}$$

Como observado do exemplo do caso 2b, a formulação de séries que não respeitem a Equação (189) não só é um fato como também pode ser originado de variações aparentemente inofensivas na configuração do problema. Resta saber as condições exigidas da formulação de um problema de RTO que assegurem a prevalência da condição fundamental (189), assim como a garantia de condições adequadas de

desempenho, como a correlação de valores apropriadamente pequenos de ε associados a valores de N também de pequena monta, de modo a restringir a oscilação a um reduzido número de ciclos.

A variabilidade induzida pelo processamento incorreto da informação, associada à cadeia de eventos apresentada na Equação (142), está diretamente vinculada ao valor e à existência de N. Para fins didáticos, de modo que as condições aqui propostas possam ser entendidas analiticamente, esta Seção irá se concentrar em exemplos apoiados nos casos 1, 2a e 2b da Seção 3.1.1.1, em que as condições determinísticas dadas pela ausência de corrupção dos sinais prestam-se à enunciação teórica dos requisitos para que as soluções consecutivas do RTO atendam a Equação (189).

Em termos pragmáticos, a atuação do RTO pode ser vista como a formulação da função prevista na Equação (190). Na ausência de variabilidade externa (**crp** = **0**, $\mathbf{Z}_j$ = $\mathbf{Z}_0$), esta atuação pode ser ainda mais simplificada, como apresentado na Equação (191), sendo expressa pela formulação de uma regra iterativa de composição de soluções. É responsabilidade da configuração do RTO assegurar que a regra de iteração φ, gerada em consequência dos procedimentos (157-160) seja coerente com a geração de uma série de Cauchy. Deste modo, a convergência estará assegurada em função das características do RTO que levam à formulação da regra de iteração φ.

$$\mathbf{Z}_{j+1}\Big|_j = \varphi(\mathbf{Z}_j) \tag{190}$$

$$\mathbf{u}_{j+1} = \varphi(\mathbf{u}_j) \tag{191}$$

Sob a ótica revelada na Equação (191), a garantia da convergência do RTO pode ser associada à solução da Equação **u** = φ(**u**). Vista desta forma, ela apresenta-se de forma análoga ao problema de descobrir os pontos fixos da função φ [60]. A partir do reconhecimento desta similaridade, resta formular a condição que rege a convergência da regra iterativa φ. Para tanto, é necessário inicialmente definir o conceito de contração [61]:

DEFINIÇÃO (Contração): *Supondo o espaço métrico X, dotado da métrica d, X=(X,d), será dito que o mapeamento T: X → X representa uma contração em X se existir um*

número real $0 \leq \alpha < 1$ tal que a relação (192) seja verdadeira. O ínfimo de α é chamado de constante de Lipschitz.

$$d(Tx,Ty) \leq \alpha d(x,y), \quad \forall x,y \in X \tag{192}$$

Isto posto, pode-se evocar o teorema do ponto fixo de Banach (ou teorema da contração) [61]:

TEOREMA 1*: Dado o espaço métrico X = (X,d), onde X ≠ ϕ, se X for completo e T for uma contração em X, então:*

- T possui um ponto fixo único em X, dado por x^*.

- para qualquer $x_0 \in X$ a sequência definida por $a_0 = x_0$ e $a_{n+1} = T(a_n)$ converge para x^*.

- a seguinte estimativa de erro a priori para a m-ésima iteração é válida:

$$d(x_m, x^*) \leq \frac{\alpha^m}{1-\alpha} d(x_0, x_1) \tag{193}$$

Portanto, uma vez que se observe o problema do RTO como um caso particular do problema genérico do ponto fixo de uma sequência obtida iterativamente, pode-se fazer uso das propriedades abstratas do teorema da contração para não só predizer a condição de convergência como também avaliar-se a velocidade de convergência, de acordo com a Equação (193). Ocorre, contudo, que frequentemente o atendimento da condição (192) com $0 \leq \alpha < 1$ não se dá em todo o espaço X, mas apenas em alguns subconjuntos seus. Isto traz como consequência a necessidade de restringir o domínio das variáveis de iteração. Como exemplo, pode-se estudar, à luz dos conceitos desta Seção, os fatos observados no caso 1 da Seção anterior. Como apresentado na Tabela 3 e na Figura 5, para este caso, dada a condição inicial $\mathbf{Zm}_0$ (Equação (153)), com $x_0 = 1$, fica clara a evolução do processo para a condição de estabilidade prevista na Equação (176). O sucesso do procedimento iterativo em atingir a condição prevista pode ser avaliado na Figura 15, onde a regra de iteração $\varphi(x)$ é dada pela Equação (175).

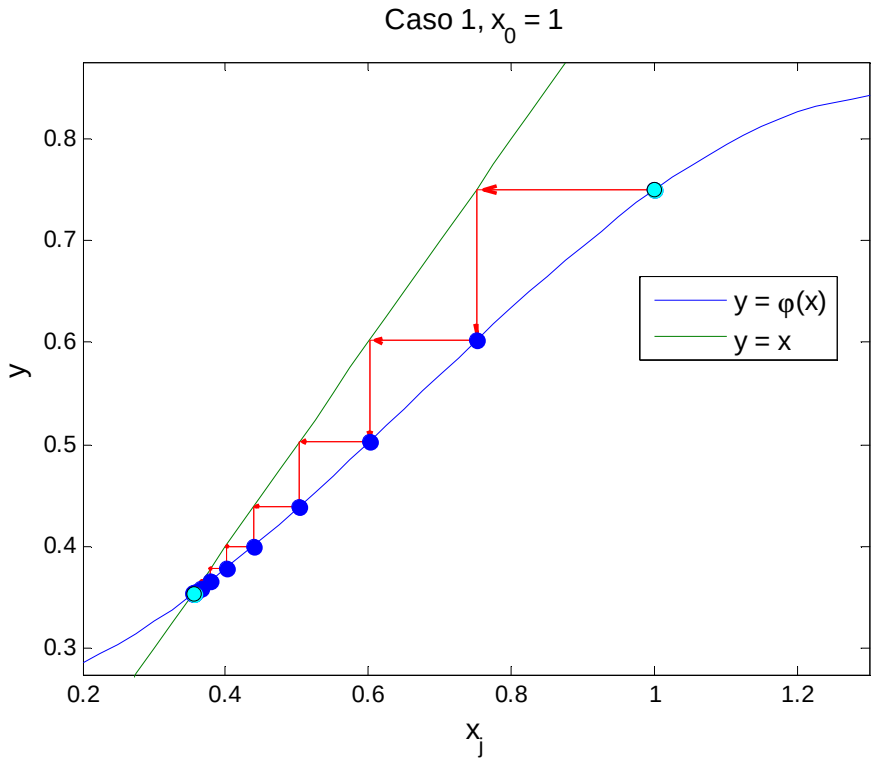


Figura 15 – Evolução do processo iterativo do RTO para o caso 1, $x_0 = 1$. $\varphi(x)$: função de iteração do RTO

As possíveis falhas do procedimento iterativo inerente ao RTO em alcançar a estabilidade prevista pela Equação (176) são evidenciadas pela análise da condição (192) ao longo do domínio o mais amplo possível. Uma forma mais conveniente de avaliação consiste em verificar o valor de α_0, como definido na Equação (194), para todos os pontos do domínio considerado. Como se pode notar a partir da Figura 16, na qual é mostrado o comportamento de α_0, existem restrições quanto ao domínio da função $\varphi(x)$, uma vez que, mesmo para o domínio restrito à faixa [-1.5 2] ocorre a violação da condição (194) em algumas regiões.

$$\left(\alpha_0 = \frac{\left\| \varphi(x_1) - \varphi(x_0) \right\|}{\left\| x_1 - x_0 \right\|} = \frac{\left\| x_2 - x_1 \right\|}{\left\| x_1 - x_0 \right\|} \right) < 1 \qquad (194)$$

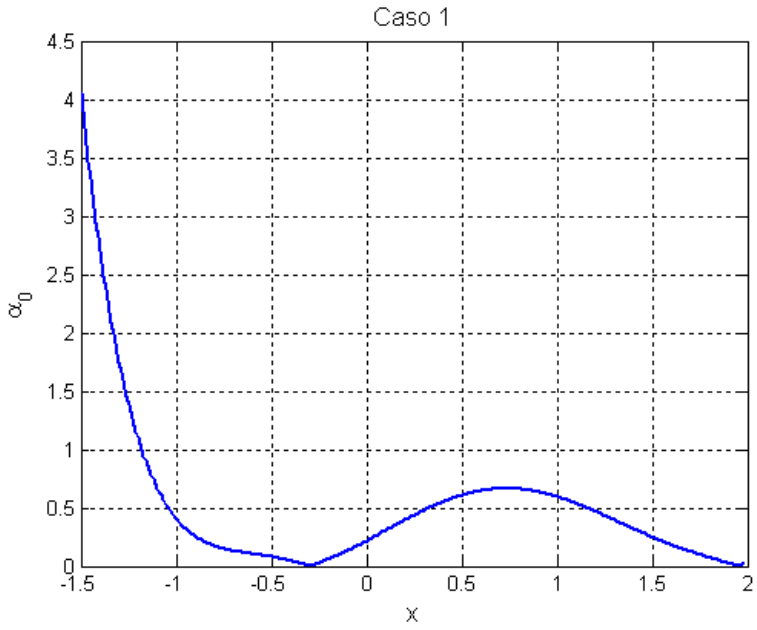


Figura 16 – Comportamento de α_0 (vide texto) para o caso 1 no domínio [-1.5 2]

De fato, como visto na Figura 16, existe uma limitação para a expansão do domínio para valores inferiores a cerca de -1,1812, valor a partir do qual α_0 torna-se maior que 1, o valor limite para a garantia da convergência. A transição da condição limite de convergência para o ponto a partir do qual o RTO propõe trajetórias não convergentes pode então ser delimitada, e estes limites devem ser ativamente considerados em suas regras de execução. O comportamento próximo deste limite pode ser observado na Figura 17, enquanto que o comportamento da variável de decisão do RTO cujo ponto de partida está ligeiramente fora deste limite pode ser visto na Figura 18. De fato, o domínio apropriado para x deve atender os seguintes requisitos: 1) conter o ponto de estabilidade; 2) garantir as condições de convergência. No presente caso, um exemplo de domínio factível seria dado por *D*: [-1 2].

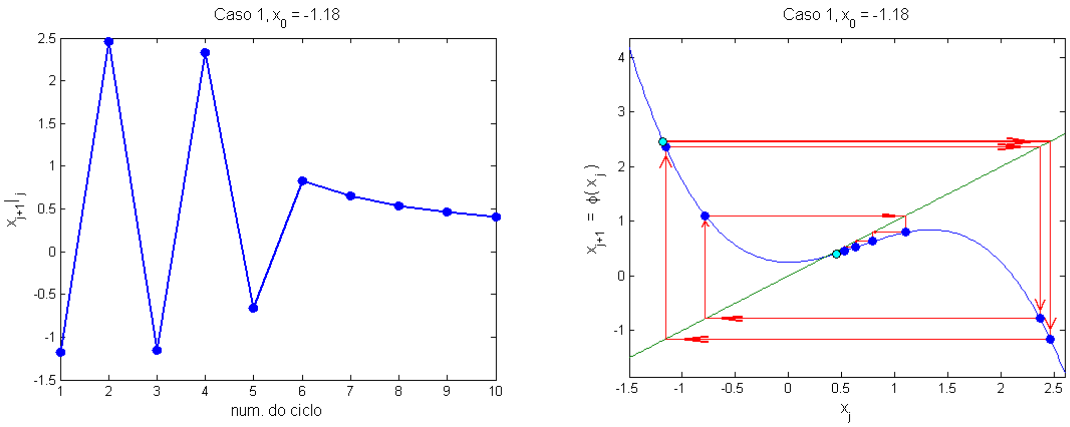


Figura 17 – Evolução de $x_{j+1}|_j$ ao longo dos ciclos (esquerda) e em função de seus elementos predecessores na iteração do RTO (direita) para $x_0 = -1.18$.

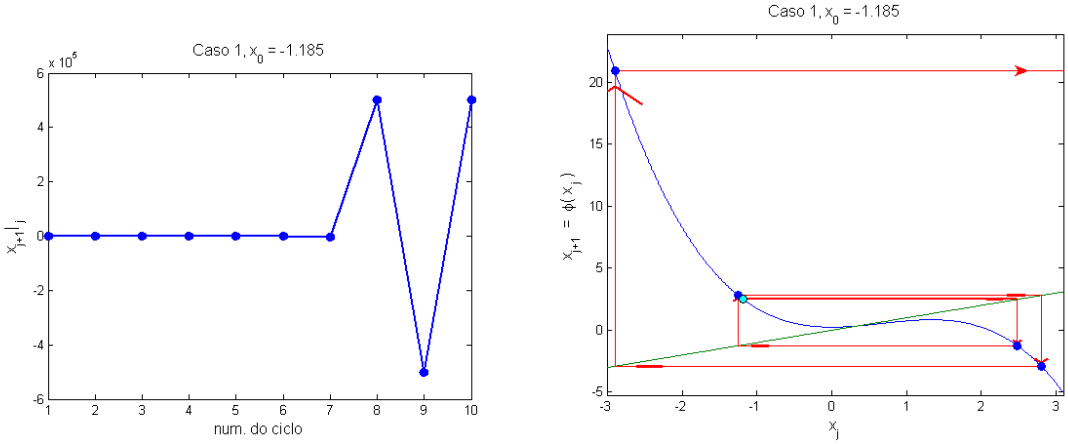


Figura 18 - Evolução de $x_{j+1}|_j$ ao longo dos ciclos (esquerda) e em função de seus elementos predecessores na iteração do RTO (direita) para $x_0 = -1.185$.

Em relação aos exemplos contidos no caso 2, torna-se menos direta a formulação analítica das funções de iteração associadas à ação do RTO, do modo como proposto para o caso 1. Contudo, é possível calcular numericamente a função φ(x) para os casos 2a e 2b por meio da conveniente manipulação algébrica e consolidação dos passos 1 a 4 (Equações 157-160), cujo resultado é apresentado na Figura 19.

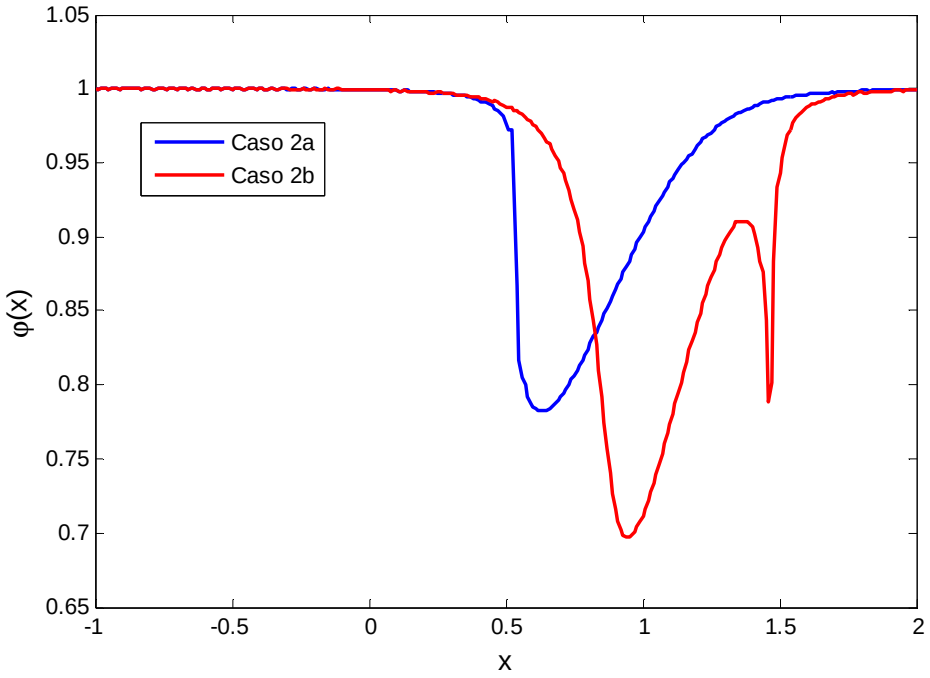


Figura 19 – Comportamento de φ(x) para o RTO dos casos 2a e 2b.

O efeito das características das funções φ(x) está diretamente relacionado com a explicação dos fatos observados na Seção 3.1.1.1 para os casos 2a e 2b. Vale lembrar que estes casos poderiam, à primeira impressão, ser considerados mais benignos, uma vez que se apóiam em modelos estruturalmente perfeitos de representação do processo, ao contrário do caso 1. Isto ficou claro para o caso 2a, através da evolução iterativa do RTO, conforme mostrado na Figura 20, que confirma a previsão de estabilidade e convergência que poderia ser feita a partir do comportamento de α_0, apresentado na Figura 21.

O comportamento aparentemente surpreendente visto no caso 2b (Figura 10), fruto da alteração em um dos elementos de $\mathbf{Z_0}$ no sentido simétrico ao experimentado no caso 2a, espelha o comportamento observado no processo iterativo cíclico resultante da ação do RTO, como visto na Figura 22.

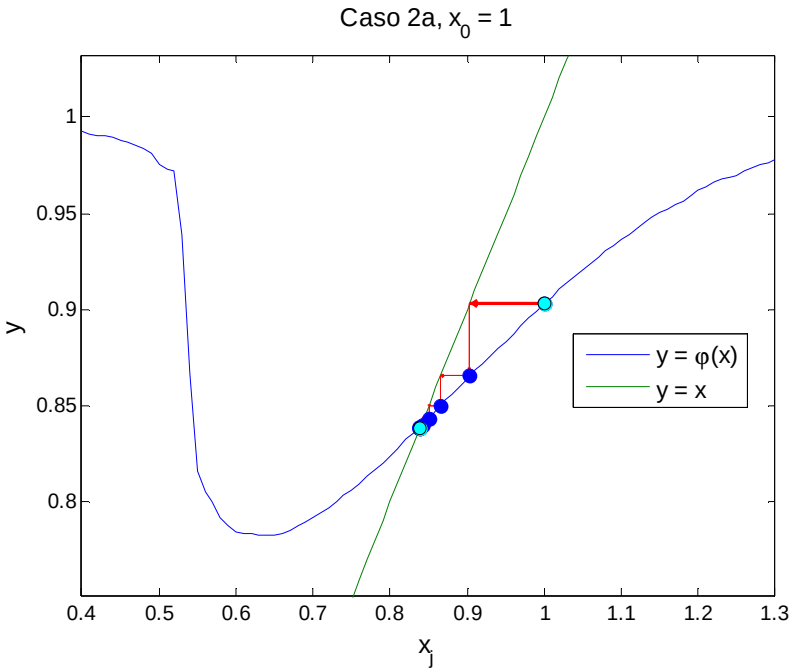


Figura 20 - Evolução de $x_{j+1}|_j$ ao longo dos ciclos e em função de seus elementos predecessores na iteração do RTO.

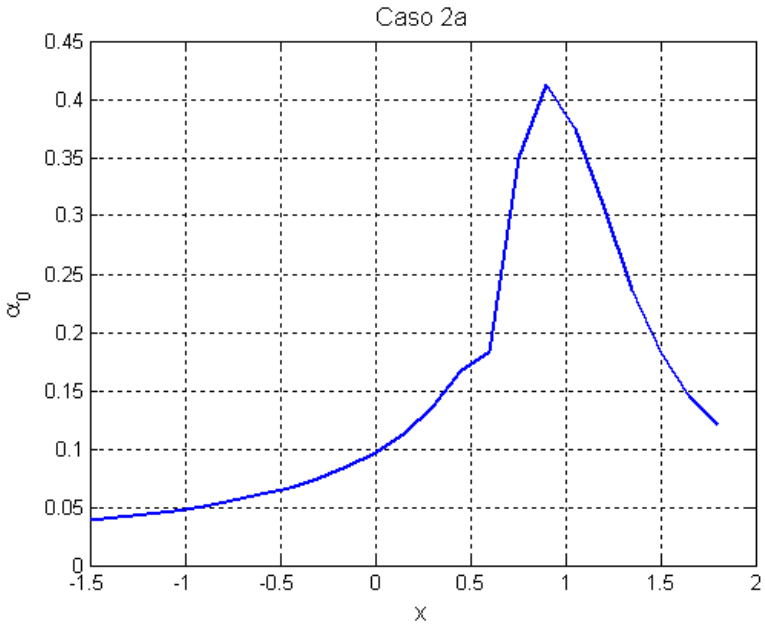


Figura 21 - Comportamento de α_0 (vide texto) para o caso 2a no domínio [-1.5 2]

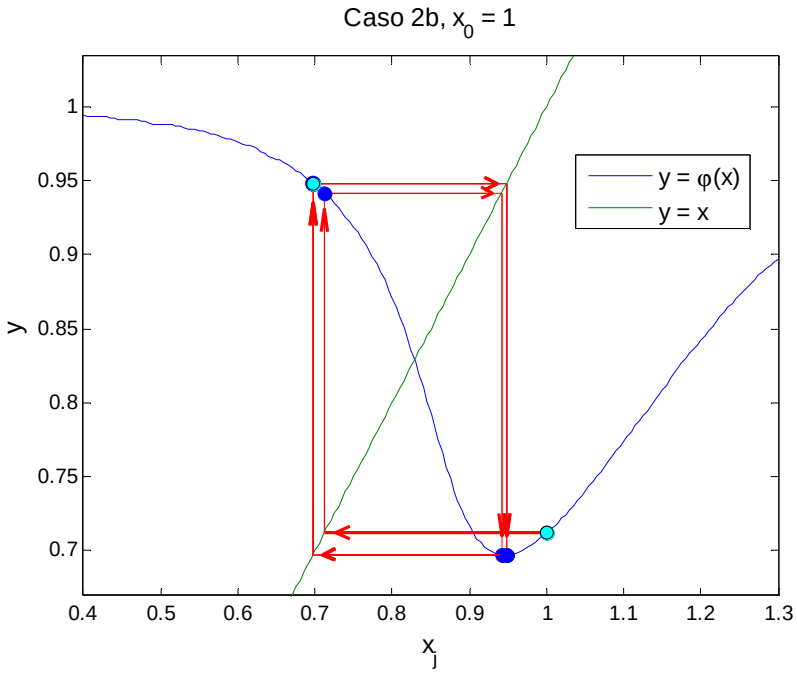


Figura 22 - Evolução de $x_{j+1|j}$ ao longo dos ciclos e em função de seus elementos predecessores na iteração do RTO.

É importante notar o ganho de generalização e abstração obtido ao se entender o problema de RTO como um processo iterativo sujeito às condições do Teorema 1. Apesar de o problema posto sob a forma $x = \varphi(x)$ ter claramente uma solução única, evidenciado pelo ponto $x \approx 0.83$ na Figura 22, o método iterativo do RTO falha em conduzir o sistema a esta solução em virtude de α_0 violar as condições previstas nas Equações (192) e (194), como constatado na Figura 23 para o domínio considerado. É importante notar a capacidade de previsão que podemos ter a respeito das possibilidades de remediação deste problema. De fato, uma vez que a solução seja a restrição do domínio considerado, a alternativa seria formulá-lo de modo a que abarcasse a solução $x = \varphi(x)$ e garantisse que a condição $\alpha_0<1$ fosse respeitada. Contudo, a Figura 23 apresenta um cenário que confirma a completa inviabilidade de se pôr em prática esta estratégia. Na verdade, a solução de $x = \varphi(x)$ infelizmente coincide com o ponto extremo de α_0. Este fato condena o RTO considerado a nunca ser capaz de, mediante o seu modo de ação iterativo, atingir a condição de estabilidade, independentemente da reformulação do domínio, a despeito do fato de a solução do problema perseguido existir.

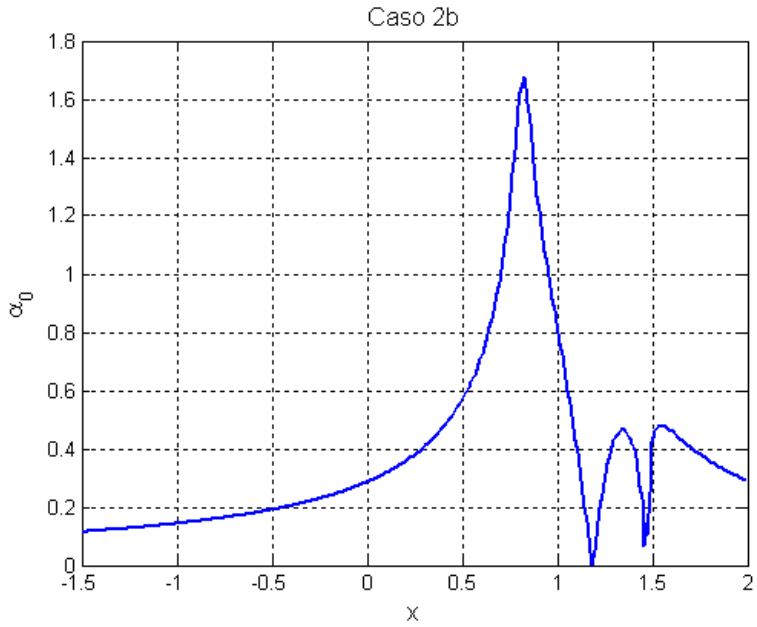


Figura 23 - Comportamento de α_0 (vide texto) para o caso 2b no domínio [-1.5 2]

3.1.2. Variações e Alternativas

Ao invés de lidar com a estrutura do modelo para garantir as condições de convergência do ponto ótimo do otimizador para o ponto ótimo real da planta, Roberts [16,62] propôs uma modificação do método em duas etapas na qual as duas camadas de otimização são desacopladas e subordinadas a um procedimento de coordenação que visa compensar falhas de predição do modelo de comportamento do processo. Neste novo procedimento, que ficou conhecido pela sigla ISOPE (*Integrated System Optimization and Parameter Estimation*), a camada de coordenação visa a garantir que o procedimento levará a planta ao ponto ótimo real ao perseguir as condições de Karush-Kuhn-Tucker [63] para o problema de otimização da camada superior, modificando a função objetivo por meio dos parâmetros de adaptação $\boldsymbol{\lambda}$, calculados como na Equação (195). O procedimento modificado pode ser resumido nas etapas descritas nas Equações (196,197). Por simplicidade, na Equação (196) as relações de igualdade foram usadas para expressar as variáveis de resposta em função das variáveis manipuladas **u** e dos parâmetros atualizados, $\boldsymbol{\theta}$. Note-se que, para suavizar a trajetória, a solução proposta $\hat{\mathbf{u}}$ é filtrada como proposto na Equação (198) para gerar o vetor efetivamente implementado $\mathbf{u}^{imp}$. Os valores da matriz diagonal **k** do filtro consistem em parâmetros de sintonia do

otimizador. Desenvolvimentos posteriores [64] do método ISOPE cuidaram da análise teórica de sua convergência e acurácia na determinação do ponto ótimo.

$$\boldsymbol{\lambda}=\left[\left[\frac{\partial \mathbf{y}}{\partial \mathbf{u}}\right]^{T}-\left[\frac{\partial \mathbf{y}_{p}}{\partial \mathbf{u}}\right]^{T}\right]\left[\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}}\right]^{-1}\left[\frac{\partial L}{\partial \boldsymbol{\theta}}\right] \tag{195}$$

onde p denota as condições reais da planta

$$\begin{aligned}&\hat{\mathbf{u}} = \min_{\mathbf{u}} \left[-L(\mathbf{u},\boldsymbol{\theta}^{+},\boldsymbol{\gamma}) - \boldsymbol{\lambda}\mathbf{u}\right]\\ &s.a.\ \ g(\mathbf{u},\boldsymbol{\theta}^{+}) \le \mathbf{0}\end{aligned} \tag{196}$$

$$\boldsymbol{\theta}^{+} = \min_{\boldsymbol{\theta}} \left(\mathbf{y}_{p} - \mathbf{y}\right)^{T} \mathbf{V}_{y}^{-1} \left(\mathbf{y}_{p} - \mathbf{y}\right) \tag{197}$$

$$\begin{aligned}&s.a\\ &\mathbf{f}(\mathbf{u},\mathbf{y},\mathbf{x},\boldsymbol{\theta}) = \mathbf{0}\\ &\mathbf{g}(\mathbf{u},\mathbf{y},\mathbf{x},\boldsymbol{\theta}) \le \mathbf{0}\\ &\boldsymbol{\theta}^{\min} < \boldsymbol{\theta} < \boldsymbol{\theta}^{\max}\end{aligned}$$

$$\mathbf{u}_{i+1}^{imp} = \mathbf{u}_{i}^{imp} + \mathbf{k}\left(\hat{\mathbf{u}}_{i+1} - \mathbf{u}_{i}^{imp}\right) \tag{198}$$

Outras propostas de mudança dos termos do problema de otimização por meio da inclusão de parâmetros aditivos foram feitas visando corrigir desvios da previsão das funções de desigualdade [17,65], desvios do valor das variáveis de resposta do processo [18] e desvios dos gradientes das funções de desigualdade [19], estes dois últimos ainda no contexto da método ISOPE.

A fundamentação teórica que garante a convergência ótima prevista esteja vinculada às condições necessárias de otimalidade (CNO) para problemas de otimização sujeitos a restrições de desigualdades, tal como previstas pela formulação de Karush-Kuhn-Tucker (KKT). Para o problema real:

$$\begin{aligned}&\mathbf{u}_{otm} = \min_{\mathbf{u}} \ -L_{p}(\mathbf{u},\mathbf{y}_{p},\boldsymbol{\gamma})\\ &s.a.\ \ \mathbf{g}_{p}(\mathbf{u},\mathbf{y}_{p}) \le \mathbf{0}\\ &\mathbf{g}_{p} = [g_{p1} \ldots g_{p,ng}]\end{aligned} \tag{199}$$

onde p denota as condições reais da planta

O objetivo é conseguir que a previsão do vetor de variáveis manipuladas, $\hat{\mathbf{u}}$, obtida através do procedimento definido na Equação (200) por meio do uso do modelo imperfeito $\mathbf{y}=\mathbf{h}(\mathbf{u},\boldsymbol{\theta})$ e $\mathbf{g}(\mathbf{u},\boldsymbol{\theta})$ seja capaz de igualar o vetor $\mathbf{u}_{otm}$, que efetivamente leva a planta ao melhor ponto de operação via (199).

$$\begin{aligned} &\hat{\mathbf{u}} = \min_{\mathbf{u}} \; -L(\mathbf{u},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &s.a. \;\; \mathbf{g}(\mathbf{u},\boldsymbol{\theta}) \le \mathbf{0} \end{aligned} \tag{200}$$

Para o problema definido na Equação (200), se existirem multiplicadores $\boldsymbol{\lambda}$ que garantam a validade das condições necessárias de otimalidade (201-204), a solução $\mathbf{u}_{otm}$ corresponderá ao valor ótimo dos graus de liberdade do problema [63,66]:

- estacionariedade do Lagrangiano:

$$\begin{aligned} &\nabla_{\mathbf{u}} \; \mathcal{L} = \mathbf{0} \\ &\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = \frac{\partial L}{\partial \mathbf{u}} + \boldsymbol{\lambda}^{\mathrm{T}} \frac{\partial \mathbf{g}}{\partial \mathbf{u}} = \mathbf{0} \end{aligned} \tag{201}$$

- condição de complementaridade

$$\boldsymbol{\lambda}^{T}\mathbf{g} = \mathbf{0} \tag{202}$$

- restrições dos multiplicadores

$$\boldsymbol{\lambda} \ge \mathbf{0} \tag{203}$$

- restrições do problema original

$$\mathbf{g} \le \mathbf{0} \tag{204}$$

As propostas [16, 17, 18, 19, 62,65] previam correções aditivas ao modelo ou às suas restrições, mantendo o procedimento de atualização dos parâmetros e sem explicitamente focar na garantia de atendimento de todas as CNO. Estas propostas evoluíram para o procedimento de adaptação de modificadores proposto por Chachuat *et al.* [13,14] e sistematizado por Marchetti *et al.* [52]. Neste caso, não há mais o procedimento de otimização em duas camadas, uma vez que os parâmetros do modelo não sofrem atualização, mantendo-se a mesma parametrização do modelo do processo ao longo da operação do RTO. Por outro lado, modificadores especificamente formulados para garantir as CNO são atualizados a cada novo estado estacionário após a implementação de uma modificação das variáveis de decisão.

O procedimento de adaptação de modificadores [13,52] propõe a atualização dos parâmetros $\boldsymbol{\Lambda} = [\boldsymbol{\delta}^{g^T}, \boldsymbol{\lambda}^{g_1\,T}, ..., \boldsymbol{\lambda}^{g_{ng}\,T}, \boldsymbol{\lambda}^{L\,T}]$, avaliados no ponto associado à k-ésima execução do RTO, e composto pelos modificadores:

- de desvio das condições de restrição:

$$\boldsymbol{\delta}^g = \left[\mathbf{g}_p(\mathbf{u}) - \mathbf{g}(\mathbf{u})\right]_{\mathbf{u}=\mathbf{u}_k}, \quad \boldsymbol{\delta}^g \in \Re^{\mathrm{ng}} \tag{205}$$

- de desvio do gradiente das condições de restrição:

$$\boldsymbol{\lambda}^{\mathbf{g}^T} = \left[\frac{\partial \mathbf{g}_p(\mathbf{u})}{\partial \mathbf{u}} - \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\right]_{\mathbf{u}=\mathbf{u}_k}, \quad \boldsymbol{\lambda}^{\mathbf{g}} \in \Re^{\mathrm{nu\ X\ ng}} \tag{206}$$

- de desvio do gradiente da função objetivo:

$$\boldsymbol{\lambda}^{L^T} = \left[\frac{\partial L_p(\mathbf{u})}{\partial \mathbf{u}} - \frac{\partial L(\mathbf{u})}{\partial \mathbf{u}}\right]_{\mathbf{u}=\mathbf{u}_k}, \quad \boldsymbol{\lambda}^{L} \in \Re^{\mathrm{nu}} \tag{207}$$

O problema de otimização modificado passa a ser expresso como:

$$\begin{aligned} \hat{\mathbf{u}} &= \min_{\mathbf{u}} \; -L(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + \boldsymbol{\lambda}^{L^T}(\mathbf{u} - \mathbf{u}_k) \\ s.a. \;& \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) + \boldsymbol{\delta}^g + \boldsymbol{\lambda}^{\mathbf{g}^T}(\mathbf{u} - \mathbf{u}_k) \leq \mathbf{0} \end{aligned} \tag{208}$$

Confrontando-se a formulação do problema (208) com as CNO, pode-se associar os modificadores $\boldsymbol{\delta}^g$ com a garantia das condições definidas na Equação (204), e os modificadores $\boldsymbol{\lambda}^L$ e $\boldsymbol{\lambda}^g$ com a garantia das condições definidas na Equação (201). Note-se que os desvios dos gradientes devem ser calculados fazendo uso de medições na planta.

Apesar da aparente similaridade com o método de atualização de modificadores, o método ISOPE permanece vinculado ao procedimento em duas etapas e propõe correções apenas na função objetivo. Na proposta de Chachuat *et al.* [13,14] as correções também se estendem aos gradientes das desigualdades e da função objetivo (206-207) abarcando de forma mais completa as CNO. Além disto, os parâmetros $\boldsymbol{\theta}$ do modelo do processo permanecem os mesmos ao longo dos ciclos de otimização, o que confere uma característica pragmática ao procedimento, na medida em que há o desacoplamento com o possível sentido físico associado à adaptação dos modificadores.

3.2. Estacionariedade de Sinais de Processo

Todos os processos que representam realidades físicas macroscópicas, em que as transferências de calor e massa se dão entre envoltórias de dimensões e massa finitas, apresentam propriedades dependentes dos estados pregressos de sua evolução temporal, o que caracteriza a impossibilidade de que variações instantâneas sejam experimentadas. A representação matemática de tais processos requer a descrição dos balanços de propriedades conservativas por meio de representações de taxas de acúmulo infinitesimais, o que é apropriadamente realizado por meio de modelos que contemplam equações diferenciais, de acordo com a moldura apresentada nas Equações (18-21). Este é o caso da totalidade dos processos químicos industriais, embora o pragmatismo possa fazer com que, sob determinadas condições de tolerância, apenas as relações algébricas (19) sejam usadas para explicar os sinais que descrevem a evolução do processo. As questões abordadas neste Capítulo dizem respeito à formulação destas condições de tolerância, assim como das consequências de seu emprego em processos químicos sujeitos à ação de sistemas de otimização em tempo real.

3.2.1. Definição do Estado Estacionário

Sob o ponto de vista estatístico, um processo cuja realização se expressa pela sequência de valores contida no conjunto {X}, de tamanho N, é dito estacionário se subconjuntos do tipo $\boldsymbol{x}(\tau,n) \equiv \{x_{\tau+1}, x_{\tau+2}, ..., x_{\tau+n}\}$ possuem função densidade de probabilidade (pdf) conjunta que seja única, independente dos valores de τ e de n. Esta definição implica que, se existirem $m_i'(\tau,n)$ e $m_i(\tau,n)$, respectivamente o i-ésimo momento e o i-ésimo momento central da pdf de $\boldsymbol{x}(\tau,n)$, vale a condição:

$$\begin{aligned} &m'_i(\tau,n) = E\left[x(\tau,n)^i\right] = \text{constante} \\ &m_i(\tau,n) = E\left[x(\tau,n)^i - \mu_{x(\tau,n)}\right] = \text{constante} \\ &\forall i,\tau,n\ ,\quad i,\tau,n \in \mathbf{N},\ \ i>0,\ \tau+n<\mathrm{N},\ \mathrm{n} \geq \mathrm{i} \end{aligned} \tag{209}$$

onde $\mu_{x(\tau,n)}$ é o primeiro momento (média) de $\boldsymbol{x}(\tau,n)$ e E[$\boldsymbol{u}$] é o operador esperança matemática de uma variável aleatória multivariável de pdf igual a $\psi(\boldsymbol{u})$:

$$E[\mathbf{u}] = \int_{-\infty}^{\infty} \mathbf{u}\,\psi(\mathbf{u})\,d(\mathbf{u}) \tag{210}$$

que, no presente caso se apresenta como:

$$E[x(\tau,n)^i] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (x_{\tau+1}x_{\tau+2}\ldots x_{\tau+n})^{i/n}\,\psi(x_{\tau+1}x_{\tau+2}\ldots x_{\tau+n})\,dx_{\tau+1}dx_{\tau+2}\ldots dx_{\tau+n} \tag{211}$$

A constância dos momentos expressa na Equação (209) caracteriza [67] a condição *stricto sensu* para a estacionariedade do processo descrito pelo conjunto {X}. Sob uma ótica menos restritiva, pode-se definir *condições fracas* de estacionariedade [67]. Para tanto, diminui-se a exigência $i \in \mathrm{N}$ em (209) tornando-a $i \in \{1,2\}$, ou seja, os dois primeiros momentos devem ser constantes para as condições formuladas na Equação (209).

Esta mudança aumenta o apelo da aplicabilidade em processos de origem física, na medida em que assume que a condição mais rigorosa é por demais idealizada. Há também um respaldo teórico amparado no fato de que, para um processo descrito por

uma distribuição normal multivariável, a constância dos dois primeiros momentos para quaisquer subconjuntos ordenados de {X} é condição suficiente para a estacionariedade, sendo equivalente à condição *stricto sensu*.

A condição fraca de estacionariedade implica que:

$$E(x(\tau,n)) = \mu \;\forall\; \tau,n \quad \text{e} \quad \text{Cov}(x_t,x_s) = \gamma_{|t-s|} \tag{212}$$

ou seja, se a seqüência {X} representa um processo estacionário de acordo com a condição expressa na Equação (212), todos os subconjuntos $x(\tau,n)$ possuem a mesma média e a covariância entre dois elementos quaisquer é função apenas de sua separação. Deve-se notar que, neste caso, o sentido da passagem do tempo, ou outra variável indexadora equivalente da sequência, é irrelevante.

Sendo verdadeiras as condições de estacionariedade previstas em (212):

$$\text{Cov}(\boldsymbol{x}(\tau,n)) = \Gamma$$

$$\Gamma = \begin{bmatrix} \gamma_{\tau,\tau} & \gamma_{\tau,\tau+1} & \gamma_{\tau,\tau+2} & \cdots & \gamma_{\tau,,\tau+n} \\ \gamma_{\tau+1,\tau} & \gamma_{\tau+1,\tau+1} & \gamma_{\tau+1,\tau+2} & \cdots & \gamma_{\tau+1,\tau+n} \\ \gamma_{\tau+2,\tau} & \gamma_{\tau+2,\tau+1} & \gamma_{\tau+2,\tau+2} & \cdots & \gamma_{\tau+2,,\tau+n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{\tau+n,\tau} & \gamma_{\tau+n,\tau+1} & \gamma_{\tau+n,\tau+2} & \cdots & \gamma_{\tau+n,\tau+n} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_0 \end{bmatrix} \tag{213}$$

para todo τ,n

As definições apresentadas nas Equações (209) e (212) são moldadas para a caracterização de estacionariedade de sinais de natureza aleatória sob o nível mais conveniente de rigor. Definições de caráter puramente arbitrário também costumam ser usadas [68], como mostrado na Equação (214)

$$\left|\frac{f(t)-f(t_0)}{t-t_0}\right| \le T_f \quad \forall\, t \in [t_0-\Delta t, t_0+\Delta t] \tag{214}$$

Neste caso, a estacionariedade é caracterizada pelo fato de o sinal apresentar taxa de variação local que não exceda determinado valor T_f para janelas de observação de comprimento $2\Delta t$ ao redor de qualquer ponto de observação. O descompromisso com o rigor estatístico e o apelo à parametrização arbitrária em prol de objetivos mais pragmáticos contidos em definições como as da Equação (214) representam a admissão de duas dificuldades reais: a de encontrar sinais que preencham o rigor de definições mais fundamentais, e de detectar, em tempo real, a presença da estacionariedade.

A classe de problemas com a qual a engenharia química e, mais particularmente, a otimização em tempo real de processos químicos se defronta é definida classicamente como consistindo de variáveis de comportamento determinístico cuja representação de sua evolução temporal é expressa por sistemas de equações diferenciais de primeira ordem do tipo $f(\mathbf{x},\mathbf{y},\dot{\mathbf{y}},\boldsymbol{\theta})=\mathbf{0}$. Sob o ponto de vista do modelo de comportamento, a estacionariedade pode ser caracterizada para intervalos em que as taxas de variação das grandezas de acúmulo sejam nulas, ou seja, $\dot{\mathbf{y}}=\mathbf{0}$.

Para lidar com sinais reais, é admitido que parte da informação oriunda do processo sob observação é de natureza aleatória, sendo acondicionada em variáveis estocásticas de caráter aditivo à previsão determinística. Esta admissão traz consigo a necessidade de discriminar, na informação disponível, se ocorrências do tipo $\dot{\mathbf{y}} \neq \mathbf{0}$ dizem respeito à não estacionariedade determinística ou se estão relacionadas à face estocástica do sinal. É ainda possível admitir-se que haja um tipo particular de estacionariedade associada a processos cíclicos em que $\dot{\mathbf{y}} \neq \mathbf{0}$ mesmo na ausência de contaminação estocástica, mas cuja natureza periódica garanta que as condições da Equação (212) sejam atendidas desde que τ seja maior que a menor freqüência de oscilação contida no sinal. Exemplo de estacionariedade deste tipo pode ser encontrada em Fath *et al.* [69].

3.2.2. Detecção do Estado Estacionário

Um sinal atenderá os requisitos de estacionariedade (209) se, em toda a sua extensão, as causas de variabilidade que o perturbem sejam mutuamente compensadas, de modo que possíveis violações dos requisitos não sejam detectadas nos limites de significância estatística. Tal permanente manutenção da estacionariedade não é comum

de ocorrer, nem por meio de uma combinação afortunada das perturbações nem por meio de mecanismos de controle de variabilidade.

O comportamento idealmente esperado para sinais oriundos da medição de processos químicos é o de contínua transição entre sucessivos estados quase estacionários com duração e freqüência inconstantes. Para fins da análise da informação oriunda do processo é necessário o uso de ferramentas que discriminem as regiões em que o sinal atende os requisitos de estacionariedade de modo a que o tratamento adequado possa ser empregado. Supondo que os dados do processo sejam obtidos de modo a obedecerem à seguinte relação:

$$f(\mathbf{x}, \mathbf{y}+\boldsymbol{\varepsilon}, \dot{\mathbf{y}}, \boldsymbol{\theta}) = \mathbf{0} \tag{215}$$

onde o modelo de comportamento determinístico é uma dada estrutura *real* $f(\mathbf{x}, \mathbf{y}, \dot{\mathbf{y}}, \boldsymbol{\theta})$, em que $\mathbf{x}$ é o vetor de variáveis independentes, $\mathbf{y}$ é o vetor de variáveis dependentes, $\dot{\mathbf{y}}$ =dy/dt e $\boldsymbol{\theta}$ é o vetor de parâmetros e constantes. A natureza estocástica da informação é expressa pelo vetor de variáveis aleatórias $\boldsymbol{\varepsilon}$. Note-se que o comportamento do sistema é observado através de ($\mathbf{y}+\boldsymbol{\varepsilon}$), que pode incorporar duas distintas causas aparentes de não estacionariedade: a variação dos momentos da função densidade de probabilidade de $\boldsymbol{\varepsilon}$ e valores não nulos de $\dot{\mathbf{y}}$. A indeterminação das origens da variabilidade torna não trivial a verificação do cumprimento das condições impostas pela Equação (209), o que parcialmente justifica o fato de a literatura técnica não apresentar um método consensual de detecção de estacionariedade.

Uma das abordagens mais simples para tratar esse problema é utilizada por Akeman [70] e Schladt [71]. Com algumas variações, o método propõe discriminar regiões de estacionariedade por meio da análise comparativa dos valores médios ao longo de duas janelas consecutivas do sinal. Supondo-se que não há garantia de que o desvio-padrão dos dados de cada janela sejam iguais ($\sigma_1 \neq \sigma_2$), a variável T_{ack}, definida na Equação (216) possuirá distribuição t de Student com n_{GL} graus de liberdade (217) se os valores contidos nas janelas i e j forem variáveis aleatórias com distribuição normal e independentes. Para dado nível de significância α pode-se testar a hipótese de os intervalos possuírem a mesma média μ, baseado nas estimativas da média, $\overline{X}$, e do desvio padrão, *s* em cada janela.

$$T_{ack} = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}} \tag{216}$$

$$n_{GL} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{(s_i^2/n_i)^2}{n_i - 1} + \frac{(s_j^2/n_j)^2}{n_j - 1}} \tag{217}$$

H0: $\mu_1 - \mu_2 = 0$. Rejeitada se $|T_{ack}| > t_{\alpha/2,nGL}$

Desta forma, para cada janela, que consiste em uma sequência de *n* pontos ordenados, podem ser estimados os limites do intervalo de confiança da média, L^{inf} e L^{sup}. Para duas janelas consecutivas, se não houver superposição dos intervalos, como expresso na Equação (218), considera-se que as médias são diferentes no nível de significância assumido e que o processo, no momento atual, não se encontra no estado estacionário. Note-se que este procedimento advém de um relaxamento adicional à condição fraca de estacionariedade.

$$L_i^{\inf} < \mu_i < L_i^{\sup}$$

$$\mu_i \neq \mu_j \text{ se } L_i^{\inf} > L_j^{\sup}, |i-j|=1 \tag{218}$$

Jubien [72] e Kim [73] apresentam métodos similares, onde a detecção do estado estacionário é condicionada à ultrapassagem de limiares da variabilidade do sinal. Para cada ponto é calculado o desvio-padrão ao longo de uma janela temporal retroativa. Para cada variável selecionada, este valor é comparado com o valor de referência do desvio padrão (σ_{ref}), obtido em condições de "reconhecida" estabilidade. Se o desvio padrão da janela móvel for menor que um limiar previamente definido ($s_i < 3\sigma_{ref}$, de acordo com [73]) é admitido que o sinal se encontra em uma região de estacionariedade. Fica a cargo do usuário do método a definição do tamanho da janela, *n*, e do limiar de estacionariedade.

Um procedimento de apelo intuitivo comumente usado [74, 75, 76] para decidir se uma região do sinal apresenta comportamento estacionário consiste em verificar se os dados pertencentes a esta região apresentam valores cuja variação evidencia uma tendência linear. Esta evidência é verificada pela existência de significância estatística no valor do coeficiente angular que ajusta o sinal a um modelo linear do tipo **x** = a**t** +b, onde o sinal **x** = $\left[x_{i-n+1}...x_i\right]^T$ está associado ao vetor temporal **t** = $\left[t_{i-n+1}...t_i\right]^T$ e representa a região do sinal de comprimento *n* cuja estacionariedade está sendo testada.

Se os erros do ajuste dos dados ao modelo possuem distribuição normal e não são correlacionados, pode-se mostrar [77] que a estatística T_0 (219) segue a distribuição t de Student com n-2 graus de liberdade:

$$T_0 = \frac{\hat{a}}{\sqrt{s^2 / S_{xx}}}, \quad S_{xx} = \mathbf{t}^{\mathbf{T}}\mathbf{t} - \frac{\left(\sum_{i=1}^{n} t_i\right)^2}{n} \tag{219}$$

onde *s* é o desvio padrão dos erros do ajuste dos dados ao modelo e $\hat{a}$ é a estimativa do coeficiente angular. Admitindo-se o nível de significância α, a hipótese H_0: a = 0 é rejeitada, ou seja, o sinal **x** não é estacionário, se a condição apresentada na Equação (220) for verdadeira:

$$\left|T_0\right| > t_{\alpha/2,n-2} \tag{220}$$

Outra forma de detecção de estado estacionário seria através da comparação da estimativa tradicional da variância, s^2, e a estimativa da variância das diferenças entre pontos consecutivos do sinal, s_d^2, conforme sugerido por Von Neumann *et al.* (1941) [78]. A detecção de diferenças significativas entre estas quantidades poderia ser usada para a detecção de tendências no sinal. Para uma janela de comprimento *n*, tem-se que:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}_i\right)^2, \qquad s_d^2 = \frac{1}{(n-1)}\sum_{i=1}^{n-1}\left(x_{i+1} - x_i\right)^2 \tag{221}$$

A variância da diferença entre dois pontos consecutivos do sinal pode ser expressa em termos da variância em cada um dos pontos através da Equação (222):

$$s_d^2 = Var(x_{i+1} - x_i) = Var(x_{i+1}) + Var(x_i) - 2Cov(x_{i+1}, x_i) \quad (222)$$

Se o sinal for estacionário, $Var(x_{i+1}) = Var(x_i)$ e se $Cov(x_{i+1}, x_i) = 0$, tem-se que:

$$Var(x_{i+1} - x_i) = 2Var(x_i) \quad (223)$$

A estatística que relaciona as duas variâncias é dada por R, Equação (224), e também expressa como C na Equação (225) [79]. R também é eventualmente expresso de forma ligeiramente alterada pela remoção do fator 2 no denominador na Equação (224), como em [78,80].

$$R = s_d^2 / (2s^2) \quad (224)$$

$$C = 1 - s_d^2 / (2s^2) \quad (225)$$

Aplicando-se a Equação (223) na expressão de R da Equação (224) nota-se que, se as condições de estacionariedade estiverem presentes, o valor esperado para a razão R tende para 1,0 para amostras grandes ($n \to \infty$) e sujeitas a variabilidade de origem estocástica, assim como C tende para 0. A dedução do cálculo dos momentos da distribuição da estatística R foi apresentada em outro artigo de Von Neumann [80]. No presente trabalho, para fins de detecção de estacionariedade, será dada preferência à variável C como base para o teste de hipóteses. Mais especificamente, será usada a variável Cs, apresentada nas Equações (226, 227), nas quais n é o número de pontos da série ordenada a ser testada. Tal preferência se dá em função do uso da Tabela com valores crítcos de Cs reportada em Young [79]. Deste modo, o sinal será considerado estacionário se o teste definido na Equação (228) for verdadeiro, onde $C_{critico}$ está associado ao nível de significância e ao número de pontos contido na janela de dados que representa o sinal.

$$C_s = C/\sigma_c \tag{226}$$

$$\sigma_c = \sqrt{\frac{n-2}{(n-1)(n+1)}} \tag{227}$$

$$Cs < C_{critico} \Rightarrow \text{sinal estacionário} \tag{228}$$

Cao e Rhinehart [81] fazem sua proposta de detecção do estado estacionário a partir da estatística R. No entanto, por razões relacionadas à carga computacional, importantes à época da escrita do seu artigo, propuseram o uso de filtros de modo a diminuir o número de operações matemáticas associadas aos cálculos da média (229), da variância tradicional (230) e da variância das diferenças entre pontos consecutivos (231). Deste modo, não é necessário especificar o tamanho da janela móvel, embora seja preciso arbitrar o valor dos três parâmetros dos filtros ($\lambda_{1..3}$) usados nos cálculos. Mais recentemente foram apresentados [82] valores críticos para análise de erros do tipo II. Este método atraiu alguns seguidores, como Bhat e Saraf [83], que propõem um procedimento heurístico para definir estes parâmetros com a finalidade de detectar variações transientes mais rapidamente sem aumentar muito a probabilidade de falsas indicações de não estacionariedade. Também foram propostas [84,85] adaptações simples do método para uso multivariável.

$$\bar{X}_{f,i} = \lambda_1 x_i + (1-\lambda_1)\bar{X}_{f,i-1} \tag{229}$$

$$s^2_{f,i} = \lambda_2\left(x_i - \bar{X}_{f,i-1}\right)^2 + (1-\lambda_2)s^2_{f,i-1} \tag{230}$$

$$s^2_{d\,f,i} = \lambda_3\left(x_i - \bar{X}_{f,i-1}\right)^2 + (1-\lambda_3)s^2_{d\,f,i-1} \tag{231}$$

Narasimham *et al.* [86] propõem um método em duas etapas para detectar estacionariedade nos dados contidos em janelas consecutivas de sinais provenientes de várias fontes. O método supõe que há estacionariedade na informação contida no interior de cada janela, o que o torna adequado apenas para a detecção de mudanças súbitas no processo, conforme reconhecido pelo próprio autor. A primeira etapa serve para auxiliar na definição do tipo de teste usado na etapa posterior. Seu objetivo é verificar se as matrizes de covariância da cada janela são estatisticamente equivalentes. A resposta a

esta questão induz o uso do apropriado teste estatístico para verificar se as duas janelas consecutivas possuem a mesma média, caso em que será assumida a estacionariedade.

Como alternativa ao método em duas etapas [86] e considerando a hipótese mais restritiva de que a matriz de covariâncias seja constante ao longo de janelas consecutivas, Narasimham, Kao e Mah [87] apresentam um teste baseado na Teoria Matemática da Evidência, de Dempster-Shafer [88]. O método também pode ser usado com sinais multivariáveis e se fundamenta na formulação de funções de credibilidade, *m(estado),* associadas a três possibilidades: 1) o processo está em estado estacionário: *m(E)*; 2) o estado do processo foi alterado de uma janela temporal para a outra *m(NE)*; 3) não é possível decidir se o estado do processo foi alterado ou não *m(I)*.

A definição das funções de credibilidade é feita de forma parcialmente arbitrária, usando valores de referência associados a níveis de significância obtidos da distribuição T^2 de Hotelling, uma vez que o teste estatístico multivariável para avaliar a igualdade das médias de dois conjuntos de dados com mesma variância segue esta distribuição.

Para agregar informações de todas as *p* variáveis que simultaneamente representam o processo, as funções de credibilidade individuais são combinadas de acordo com as regras multiplicativas da Equação (232). A decisão sobre a continuidade do estado estacionário ao longo das janelas é tomada se o valor combinado para todos os sinais do processo e das funções de credibilidade para o estado estacionário for maior que o valor correspondente para a não estacionariedade, ou seja, se *m(E)>m(NE)* na Equação (232).

$$
\begin{aligned}
m(E) &= \prod_{i=1}^{p} \left[m_i(E) + m_i(I) \right] \\
m(NE) &= \prod_{i=1}^{p} \left[m_i(NE) + m_i(I) \right] \\
m(I) &= \prod_{i=1}^{p} m_i(I)
\end{aligned}
\qquad (232)
$$

Todos os métodos até aqui apresentados nesta Seção partem de hipóteses acerca da natureza da distribuição das funções densidade de probabilidade das variáveis estudadas. Tal hipótese é, invariavelmente, a da presença de distribuições gaussianas. É incomum, na literatura de engenharia, o uso de testes não paramétricos, ou seja, que não requeiram tais hipóteses acerca da distribuição, para identificar regiões de

estacionariedade. Önöz e Bayazit [76] se propuseram a comparar o desempenho do tradicional teste de significância estatística do coeficiente angular (219) com o teste não paramétrico de Mann-Kendall para análise de tendências.

Neste teste, a estatística *S* na Equação (233) calcula, em cada janela de *n* pontos do sinal **x**, o valor acumulado $S = P\text{-}M$, onde *P* é o número de pares em que $x_j > x_k$, e *M* é o número de pares em que $x_j < x_k$, para $j > k$, e $(j\text{-}k) < n$.

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sinal\left(x_j - x_k\right)$$

$$sinal(x_j - x_k) = \begin{cases} 1, & se\ x_j - x_k > 0 \\ 0, & se\ x_j - x_k = 0 \\ -1, & se\ x_j - x_k < 0 \end{cases} \tag{233}$$

Para séries de comprimento $n > 10$ e na ausência de valores repetidos, a estatística normalizada *Z* (Equação 234) segue a distribuição normal com média zero e variância unitária.

$$Z = \begin{cases} (S-1)/\sigma_s, & se\ S > 0 \\ 0, & se\ S = 0 \\ (S+1)/\sigma_s, & se\ S < 0 \end{cases} \tag{234}$$

$$\sigma_s = \sqrt{\frac{n(n-1)(2n+5)}{18}}$$

Os desempenhos do teste paramétrico e do não paramétrico foram comparados [76] em termos da probabilidade de ocorrência de erros tipo II por meio de simulações de Monte Carlo com séries de 2000 pontos e janelas de n = 50 valores. Foram geradas séries seguindo cinco diferentes funções densidade de probabilidade. O desempenho do teste não paramétrico foi superior para distribuições não normais e assimétricas, sendo similar para os demais casos, embora deva-se ressaltar que o teste de Mann-Kendall pode não ser eficaz para alterações não monotônicas do estado estacionário [89].

Diversos outros testes não paramétricos têm o potencial de serem utilizados para a detecção de estacionariedade. Podem ser citados o teste de Siegel-Tukey [90], o de Wilcoxon [90], o de Wald-Wolfowitz [91], o de Spearman [91], o de Friedman [92] e o *runs test* [92], muito embora este potencial não tenha sido explorado, até onde vai o

conhecimento deste autor, pela literatura técnica de engenharia química para detecção de estado estacionário.

Uma abordagem distinta das até aqui apresentadas faz uso de informações obtidas pela transformada wavelet do sinal como subsídio para a detecção de regiões de estacionariedade [68,93,94,95,96]. De forma similar à transformada de Fourier, a transformada wavelet projeta uma representação do sinal em uma base ortogonal no domínio da freqüência. É premissa da representação no domínio da freqüência a partir da transformada de Fourier que o sinal seja estacionário. A informação obtida não está associada a uma localização temporal definida, ou seja, não é possível distinguir de que forma a energia contida em determinada janela de freqüências varia ao longo do tempo. Por outro lado, o suporte restrito das funções da base de wavelet está associado a uma representação que possui uma janela de resolução no tempo e na freqüência de dimensões finitas e adaptáveis, que permite acompanhar a dinâmica das variações experimentadas pelo sinal bem como extrair padrões de comportamento em diversas escalas de resolução.

Esta versatilidade pode ser aproveitada para representar trechos do sinal por meio de aproximações polinomiais das funções que formam a base da representação de wavelet. A estacionariedade pode ser representada como um caso particular, em que o melhor ajuste está associado a um polinômio de ordem zero [93]. Jiang *et al.* [68] propõem o uso de wavelets para o condicionamento do sinal (remoção de anormalidades e perturbações não gaussianas) e identificação do grau de estacionariedade para cada instante do tempo em sistemas multivariáveis. Para agregar os índices de estacionariedade individuais associados a cada sinal oriundo do processo, Jiang *et al.* [68] propõem uma regra de combinação multiplicativa similar à usada por Narasimham, Kao e Mah [87] e ponderada por valores arbitrários associados à importância de cada sinal. Caumo e Trierweiler [96] propõem uma extensão deste método usando análise de componentes principais para agregar as informações multivariáveis para a decisão de estacionariedade.

Nesta seção, que fez uma retrospectiva deste tema na literatura técnica, os métodos prevêem abordagens restritas à morfologia dos sinais, de modo desconectado com o processo associado. Na Seção seguinte esta questão será discutida mais detalhadamente.

3.2.3. Considerações e análise crítica

Um tema não abordado explicitamente nos métodos de detecção de estado estacionário diz respeito à análise constitutiva dos sinais. A abordagem geral considera que parte da informação oriunda da medição da variável de processo é irrelevante para a finalidade última, seja esta a otimização da função econômica, seja a adaptação do modelo. A demarcação da fronteira entre a significância e a irrelevância, contudo, fica muitas vezes oculta sob hipóteses simplificadoras muito restritivas.

O uso da Equação (215) em um modelo determinístico de comportamento de um processo estacionário supõe que:

1) $\dot{\mathbf{y}}=\mathbf{0}$, o que determina a estacionariedade em si, ou seja, que o processo pode ser completamente descrito a partir do conhecimento de seu estado no momento atual;

2) $\boldsymbol{\varepsilon} = \mathbf{0}$, de modo que o mapa $\mathbf{u} \rightarrow \mathbf{y}$ dado pela estrutura do processo sempre relacione corretamente as informações, havendo completa equivalencia entre os valores medidos e os efeitos induzidos por **u**.

A violação da condição 1 impõe a necessidade de se incorporar o conhecimento dos estados anteriores do processo ao modelo descritivo, o que, à primeira vista, depende tão somente de que suficiente esforço seja despendido na modelagem do processo. Uma violação de 1 poderia, em primeira análise, ser superada mediante uma decisão que está sob o comando do desenvolvedor.

A violação da condição 2 pode ser superada se for possível representar $\boldsymbol{\varepsilon}$ por relações funcionais dependentes de grandezas conhecidas, o que tornaria este caso idêntico ao anterior. Contudo, isto conflita com a hipótese comumente assumida de que as ocorrências de $\boldsymbol{\varepsilon}$ são eventos aleatórios oriundos de um espaço amostral de realizações e mediados por dado nível de probabilidade de ocorrência. Outro modo de superar esta violação seria por meio da medição direta de $\boldsymbol{\varepsilon}$, o que esbarra em limitações práticas.

Para este segundo caso, ainda que não haja a promessa (mesmo que idealizada) de que os danos causados à capacidade preditiva do modelo possam ser superados mediante suficiente esforço, há a perspectiva de que a repetida execução de eventos em

que **ε** se manifeste aumente indiretamente o conhecimento do valor verdadeiro de **y**, conforme mostrado na Seção 3.3. A dificuldade reside no fato de que, na presença de ambas as violações, este procedimento não conduz a um valor *verdadeiro* para uma janela temporal. Tal conceito não mais se aplica uma vez que há um valor verdadeiro para toda a janela temporal de dados.

Infelizmente, ambas as violações descritas são esperadas em sinais reais de oriundos processos químicos. A estratégia clássica de lidar com este caso consiste no seguinte método seqüencial:

1) decidir se d**Z**(**ee**)/dt = **0**, onde **ee** pertence ao conjunto das variáveis medidas que foram escolhidas para a verificação de estado estacionário sob dado nível de tolerância (por meio de qualquer dos métodos descritos na Seção 3.2.2);

2) caso o estado estacionário seja detectado, seguir com os procedimentos de adaptação e otimização na suposição de erros aleatórios aditivos. Note-se que há três decisões importantes contidas na etapa 1: as definições do conjunto **ee**, do método de detecção e da tolerância assumida.

É importante assinalar que estas decisões ajudam a moldar a magnitude e a geometria da região de incerteza dos erros no espaço das variáveis de decisão, com conseqüência na estimativa do lucro esperado para a operação, embora este fato não seja explorado na literatura técnica da área de RTO.

Dentre os diversos métodos de detecção de estado estacionário, existem alguns pontos que devem ser ressaltados:

1) Há pouca comparação entre os métodos e nenhuma quantificação de seus benefícios para a operação do processo. O modo de medir o desempenho é auto-validado, na medida em que esta medição é amparada na definição de estacionariedade que está implícita na construção do método. Assim, um método de verificação da variação das médias entre janelas temporais [70,71] é considerado bom para a detecção de estado estacionário se conseguir indicar bem se as médias variam (ou não) ao longo das janelas. Assim como ocorre para os que buscam uma tendência linear no sinal [74, 75, 76], e assim sucessivamente para os demais métodos apresentados. Mesmo as comparações são dificultadas pela carência de um padrão objetivo. Por exemplo, há uma

comparação [68] entre o método de *wavelet* e o de Narasimhan [87], porém ela fica restrita a ser meramente expositiva ou subjetivamente direcionada aos critérios do método de *wavelet*.

Além disto, na ausência de critérios objetivos para referenciar um *padrão-ouro* de estacionariedade que permita identificá-la em trechos de sinais experimentais, recorre-se comumente à subjetividade para validá-la.

2) Os métodos são monovariáveis, apesar da natureza da totalidade dos processos químicos não o ser. As análises são aplicadas a cada um dos sinais, individualmente, e a conclusão final é tomada por meio de alguma regra lógica implementada pelo usuário, como por exemplo: se o número de sinais considerados como estacionários é maior que dado valor, então o processo está estacionário. Este tipo de estratégia ignora o fato de que o nível de tolerância pode ter de ser diferente para cada sinal de modo a manter restrito o impacto sobre o processo. Além disto, diferentes combinações da mesma variabilidade total distribuída ao longo dos sinais podem repercutir no processo de modo diverso. Os poucos métodos que propõem algum tipo de uso multivariável [68,84,87,94] o fazem segundo regras de combinação das análises que são exteriores à natureza do processo.

3) Embora a planta disponha de muitas variáveis medidas, apenas um subconjunto é usado para fins de detecção do estado estacionário. O método de escolha deste grupo é variado, embora sempre qualitativo. Algumas vezes é escolhido de forma arbitrária, sem explicação formal [73], em outras é apresentada como justificativa a importância de algumas variáveis para o comportamento dinâmico [71], embora sem comprovação formal deste fato. Também é apresentado como justificativa a eleição de um conjunto de variáveis que possuam correlação [83,86], embora o motivo oposto também seja usado [68]. Em ambas as linhas de ação nenhuma quantificação de correlação é apresentada.

Não é ocioso lembrar que todas estas questões poderiam ser unificadas se fossem observadas sob a ótica de sua *utilidade* para o RTO, no sentido de quantificar a diminuição da incerteza do valor ótimo previsto pelo otimizador para a função objetivo, como será visto na Seção 3.2.4. Contudo, todos os métodos de detecção de estado estacionário disponíveis na literatura foram concebidos sob a perspectiva do sinal em si

mesmo, de sua variabilidade e morfologia, porém de forma desconectada com o processo.

A conseqüência desta desconexão é que não há como determinar *a priori* qual a relação entre o nível de confiança estatística usado na parametrização do método e a incerteza associada à estimativa do lucro ótimo operacional. Ao longo das subseções 3.2.3.1 e 3.2.3.2 será mostrado como os diferentes conceitos de estacionariedade influenciam nas conclusões apontadas por cada método e o descasamento entre as detecções e os objetivos do RTO.

3.2.3.1. Desempenho comparativo de testes de estacionariedade

Como discutido na Seção anterior, cada método de detecção de estacionariedade está vinculado à verificação da morfologia sob a ótica de uma definição particular de estacionariedade. A multiplicidade de definições pode dar origem à multiplicidade de veredictos, e este aspecto será ressaltado na presente Seção. Para fins de comparação, serão avaliados os seguintes testes de detecção de estado estacionário:

- parâmetro R, de von Neumann, que testa a razão entre a variância tradicional e a variância calculada a partir da diferença de estados consecutivos, de acordo com as Equações (221-224) . Será indicado pela legenda *R*.

- significância do coeficiente angular do ajuste linear dos dados, de acordo com as Equações (219-220). Será indicado pela legenda H_a.

- Ackeman e Schladt, que testa a hipótese de a média de duas janelas consecutivas ser igual, de acordo com as Equações (216,218). Será indicado pela legenda *ack*.

- Mann-Kendall, não paramétrico, que testa a hipótese de existência de tendência monotônica em uma série baseada na variabilidade conjunta dos dados disponíveis, de acordo com as Equações (233,234). Será indicado pela legenda *mank*.

- Kolmogorov-Smirnov [97], não paramétrico, que testa a hipótese de que dois conjuntos de dados sejam amostras de uma mesma população e compartilhem a mesma função distribuição de probabilidade. Será indicado pela legenda *KSmir*.

Os cinco testes serão usados para avaliar relações funcionais do tipo apresentado na Equação (235), onde uma função determinística, $f(\mathbf{x})$, é contaminada com um vetor amostrado de uma distribuição gaussiana, $\boldsymbol{\varepsilon} \sim N(0,\sigma_\varepsilon)$, sendo que $\mathbf{x}$ pertence ao domínio [0 1] e é regularmente amostrado com período 1/100. Para os testes de Ackemann-Schladt e de Kolmogorov-Smirnov o vetor x foi dividido em duas janelas consecutivas pertencentes, respectivamente, aos domínios [0 0,5] e [0,5 1]. O nível de significância, α, assumido para todos os testes, foi de 5%.

$$\begin{aligned} &\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}; \\ &f(\mathbf{x}) \in \{ax; sen(2\pi x/T); \cos(2\pi x/T); ax + A\cos(2\pi x/T)\} \end{aligned} \tag{235}$$

O primeiro exemplo para a aplicação dos testes de estacionariedade faz uso da função linear f(x) = ax onde $a \in$ [0 0,2], sendo o desvio padrão do ruído caracterizado por $\sigma_\varepsilon \in$ [0,004 0,1]. Como pode ser visto na Figura 24, as regiões no plano a-σ_ε são identificadas com níveis de probabilidades similares para os diversos métodos, havendo, contudo, tendência a maior chance de indicação de estacionariedade no método da razão R. Ao lado deste fato, pode-se perceber que o dilema contido na formulação apresentada na Equação (215) é bem evidente: o de que a presença combinada de taxa de variação e ruído não nulos interfere diretamente nas premissas do método de identificação e, consequentemente, na expectativa de diagnóstico. Pode-se observar melhor este fato na Figura 25, onde a probabilidade de diagnóstico positivo para a estacionariedade é apresentada em função do desvio padrão de ε para o teste da razão R e para o teste de significância do coeficiente de inclinação.

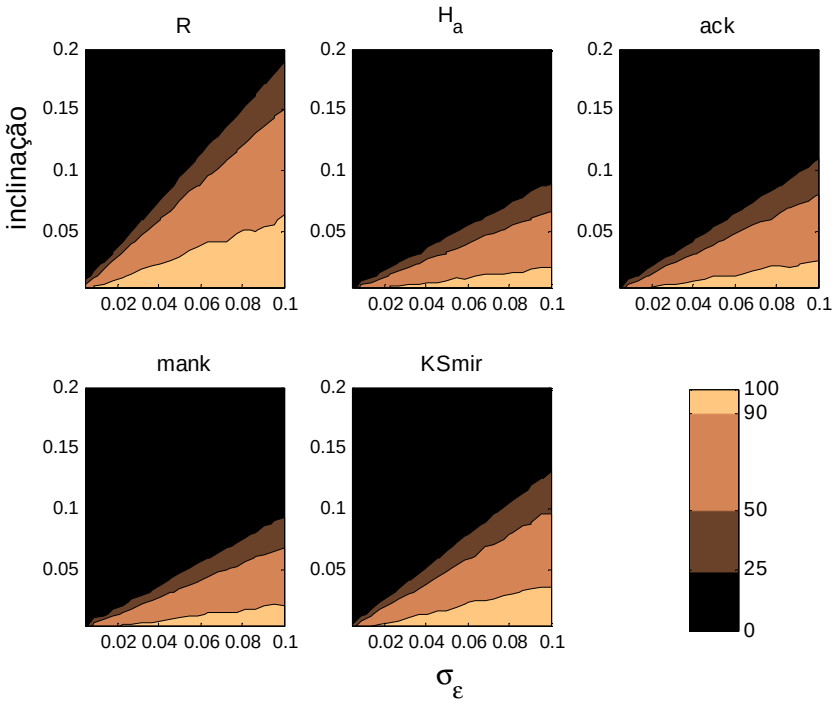


Figura 24 – Probabilidade (%) de indicações de estacionariedade de diversos testes aplicados em f(x) = ax+ε em função do coeficiente angular da função de teste, e do desvio padrão do ruído.

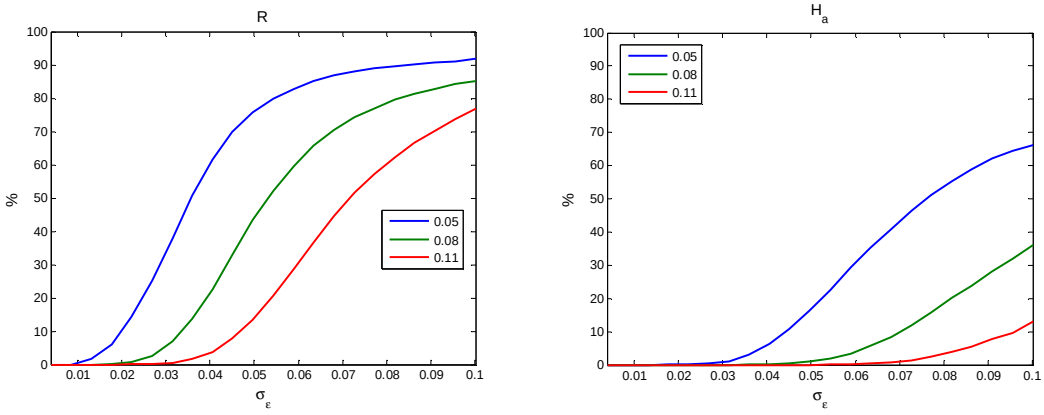


Figura 25 – Probabilidade (%) de diagnósticos de estacionariedade aplicados em f(x) = ax+ε em função do desvio padrão do ruído para dois testes de estacionariedade. Legenda: valores do coeficiente angular de f(x).

A diferença entre as respostas dos diversos métodos pode ser significativamente aumentada em virtude do tipo de função submetida à análise. Para o caso em que f(x) seja uma função cossenoidal nos moldes apresentados na Equação (235), as diferenças nas expectativas de indicação de estacionariedade dos diversos métodos tornam-se mais pronunciadas, o que pode ser visto na Figura 26. Embora a maior parte dos testes difira entre si por diferenças graduais na forma das regiões de expectativa, é de se ressaltar

que, neste caso, o teste da razão R não indica qualquer indício de estacionariedade em todos os pontos da região explorada. Pode-se também notar que, para alguns valores do período do cosseno, conclusões opostas podem ser inferidas da aplicação dos testes, como evidenciado na Figura 27. Para T = 0,67, enquanto alguns testes (H_a e *mank*) indicam a presença de comportamento estacionário com elevada probabilidade, independentemente da intensidade do ruído, outros testes (*ack* e *KSmir*) praticamente não detectam a estacionariedade. Por outro lado, esta tendência é praticamente invertida para um valor do período de 1,08, exceto para *ack*.

Não só o período, mas também a fase do sinal podem influenciar significativamente a indicação de estacionariedade dos diversos métodos. Se a função de teste for senoidal, nos moldes da Equação (235), mudanças substanciais na probabilidade de indicação de estacionariedade podem ser induzidas pela variação de 90° na fase da função periódica. No presente caso, notadamente nas regiões em que T<1 e T ≈ 2 estas alterações são mais pronunciadas, como percebido da comparação das Figuras 26 e 28.

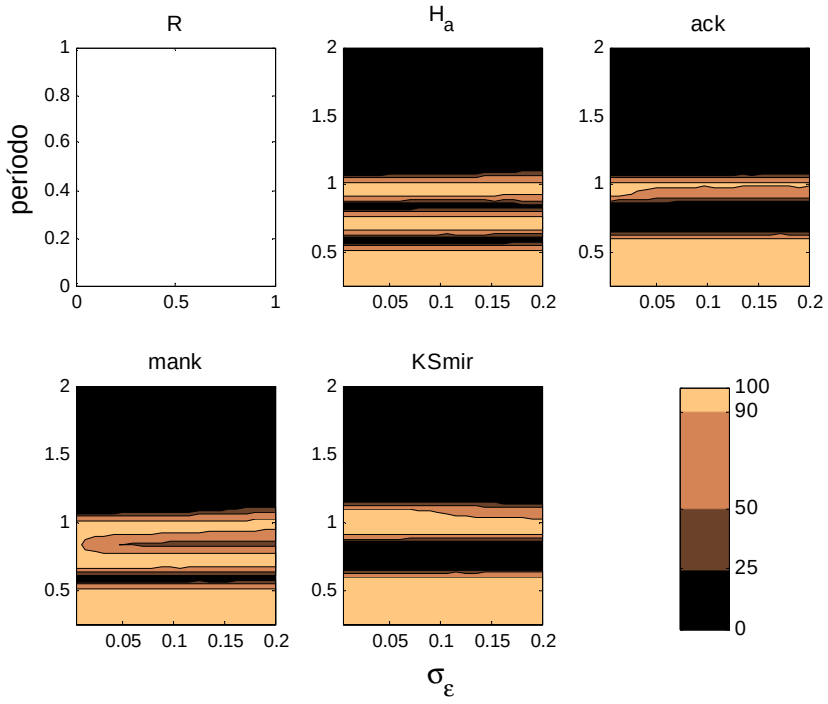


Figura 26 – Probabilidade (%) de diagnósticos de estacionariedade indicados por diversos testes para função cosseno em função do período e do desvio padrão do ruído. A cor branca no gráfico de R indica um valor de 0% de probabilidade de indicação de estacionariedade.

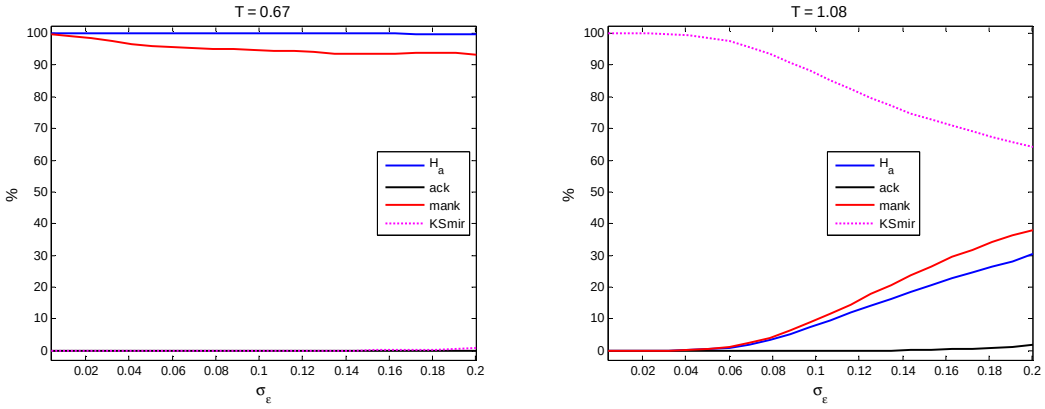


Figura 27 – Percentual de diagnósticos de estacionariedade para função cosseno em função do desvio padrão do ruído para diversos testes de estacionariedade. Os dois gráficos diferem pelo período do cosseno em f(x), indicado nos títulos.

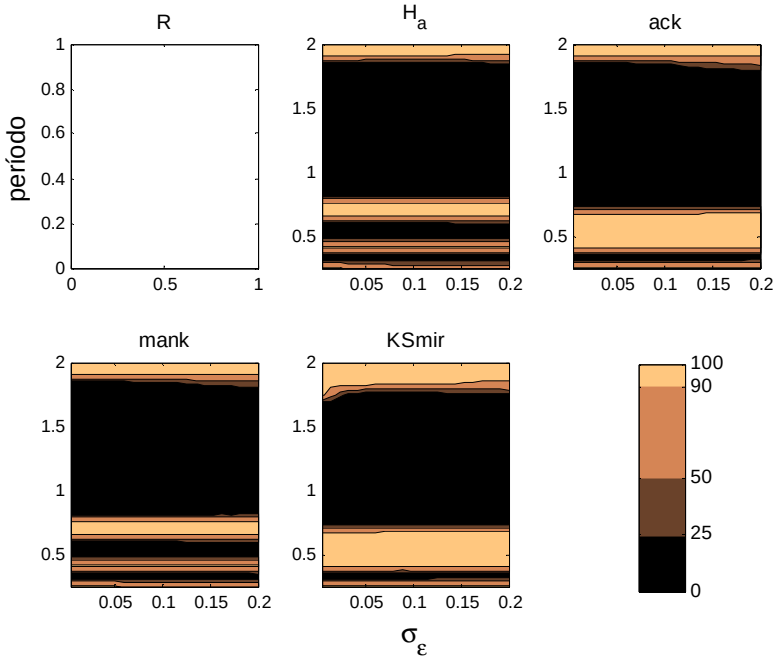


Figura 28 – Probabilidade (%) de diagnósticos de estacionariedade indicados por diversos testes para função seno em função do período e do desvio padrão do ruído. A cor branca no gráfico de R indica um valor de 0% de probabilidade de indicação de estacionariedade.

Outro exemplo interessante de ser explorado é a influência que a adição de um *trend* linear tem sobre a conclusão acerca da estacionariedade de um sinal periódico quando visto sob a ótica de detecção proposta pelos diversos métodos. Se um sinal combinado, em que f(x) = $ax + A\cos(2\pi x / T)$, como proposto em (235), for examinado sob o ponto de vista da estacionariedade, conclusões praticamente opostas podem surgir

como resultado da influência da adição do *trend* linear *ax* à função periódica (com A=0,25). Note-se, do exame das Figuras 29 e 30, que a probabilidades de detecção de estacionariedade em função do *trend* adicionado alternam-se de forma quase complementar quando comparada a função cosseno com dois períodos distintos. Enquanto que, para T = 0,25 a estacionariedade é detectada com maior probabilidade com *trends* aditivos de inclinação na faixa inferior do range, para T = 1,08 a estacionariedade é mais frequentemente detectada na faixa superior do *range* de inclinações. Vale deixar registrado que, como se esperaria, expandindo-se o *range* para abarcar inclinações para além de a=0,2 a tendência em ambos casos é a de decréscimo na probabilidade de indicações de estacionariedade.

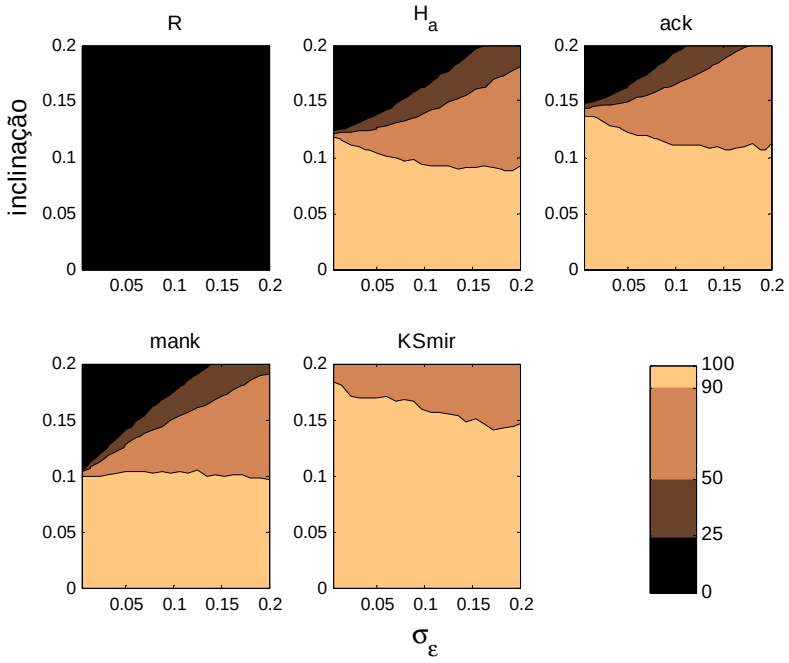


Figura 29 – Probabilidade (%) de diagnósticos de estacionariedade indicados por diversos testes para função cosseno (período = 0,25) com *trend* linear em função da inclinação e do desvio padrão do ruído.

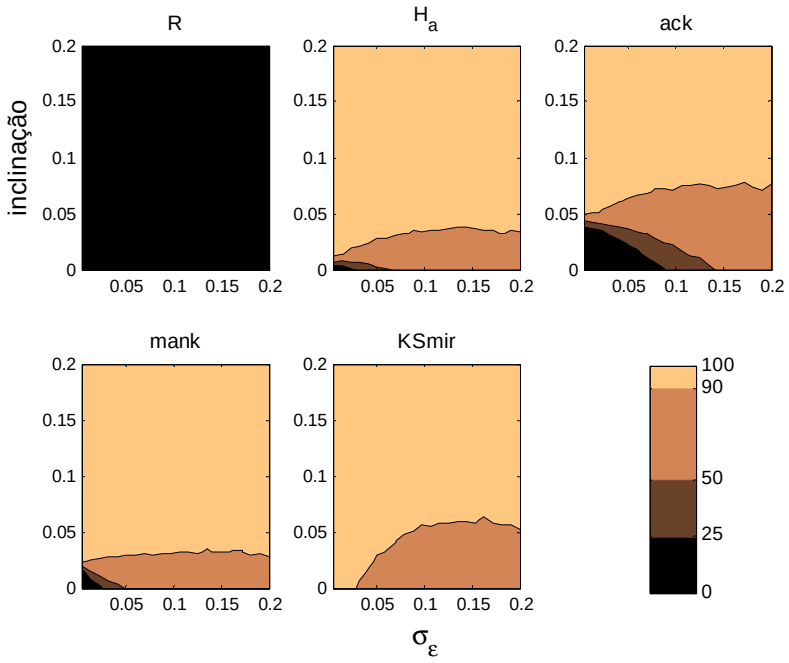


Figura 30 – Probabilidade (%) de diagnósticos de estacionariedade indicados por diversos testes para função cosseno (período = 1,08) com *trend* linear em função da inclinação e do desvio padrão do ruído.

Os casos formulados a partir da moldura contida na Equação (235) e cujas características foram apresentadas nas Figuras 24-30 exemplificam: 1) a variabilidade das conclusões acerca da estacionariedade esperada para os diversos métodos; 2) o caráter probabilístico destas conclusões; 3) a mudança não intuitiva das expectativas das conclusões em função de mudanças na parametrização das funções teste.

Com base nas exemplificações apresentadas, pode-se perceber indícios que comprometem o uso deste tipo de abordagem, que é de cunho genérico, no contexto de aplicações de RTO, pois cada método está comprometido com as condições previamente definidas em sua formulação. Deste modo, algumas questões relevantes merecem ser destacadas:

- o método de detecção da significância da inclinação, Ha, baseia-se em um teste de hipóteses estatístico fundamentado na significância do coeficiente angular de um modelo polinomial de primeira ordem *ajustado* aos dados. Quanto a isto, merece destaque o seguinte questionamento: qual o significado do teste deste coeficiente quando realizado para formas de onda distintas da proposição inicial, como as funções de caráter periódico apresentadas nas Figuras 26-30? O teste está intimamente relacionado à

suposição de contaminação gaussiana, e esta influência afeta diretamente a habilidade de discriminação da detecção da estacionariedade. Como visto nas Figuras 24 e 25, inclinações pequenas só dão origem a maior probabilidade de veredictos de estacionariedade quando associadas a ruído mais intenso. Contudo, a definição comparativa de "pequenas" e "mais intenso" na frase anterior está associada a características intrínsecas do método de detecção, e não à repercussão no desempenho do RTO, o que permite a re-edição do questionamento acerca da *utilidade* do método.

- o método de Ackeman-Schladt padece de problemas similares ao teste de significância do coeficiente angular ajustado. Assim como o teste Ha, este também é um teste que se baseia em um considerável apelo intuitivo. Descrevendo-o de forma sucinta, a idéia é que, para dado sinal $\boldsymbol{x}(\tau,n) \equiv \{x_{\tau+1}, x_{\tau+2},...,x_{\tau+n}\}$, a estacionariedade estaria configurada se a hipótese $\mu(x(\tau,n/2))= \mu(x(n/2+1,n))$ não fosse refutada, ou seja, se não se pudesse afirmar que a média dos valores da primeira e segunda metades do sinal fossem diferentes. Note-se que, olhando-se com rigor, a informação contida em $x(\tau,n)$ é de natureza estocástica, o que implica a existência de uma pdf multivariável $\psi(\mathbf{x}(\tau,n))$ a partir da qual o sinal contido em cada janela de comprimento n seja amostrado, sendo que seus momentos (incluindo a média) devem ser calculados de acordo com a Equação (211). Contudo, ao supor que a proposta formulada nesta equação pode ser substituída pela média simples dos valores de cada janela (236) implica que se assume, a priori, duas hipóses ocultas: a de que os valores de cada instante no tempo são variáveis aleatórias independentes entre si (237) e de que a variância é constante ao longo da sequência (238). Como feito comumente nestes tipos de métodos, tais hipóteses não são sujeitas a posterior validação.

$$\mu(x_1) = \mu(x_2) = ... = \mu(x_J) = \sum_{i=1}^{J} \frac{x_i}{J} \tag{236}$$

$$\psi(x_1, x_2,...,x_J) = \psi(x_1)\psi(x_2)..\psi(x_J) \tag{237}$$

$$\sigma_1 = \sigma_2 = ... = \sigma_J \tag{238}$$

Além do uso não verificável das condições (236-238), há uma presunção extremamente restritiva no método, ao menos para fins de aplicações em RTO. Uma vez que a janela de dados é arbitrariamente dividida em seu ponto médio há a predisposição de que apenas variações súbitas ao redor da posição central sejam associadas à não estacionariedade, ou, de forma mais genérica, os veredictos são mais associados à existência de assimetrias na forma de onda do que à presença de não estacionariedade em si. Esta afirmação pode ser melhor entendida observando-se os gráficos da Figura 31, onde dois sinais sofrem variações súbitas de mesma intensidade, porém defasadas. No caso em que a transição ocorre no ponto em que as sub-janelas de comparação são divididas, há a indicação de não estacionariedade. No caso em que a transição ocorre em outro ponto, de modo que as médias dos sinais em ambas as janelas coincidam, há o veredito de estacionariedade, ainda que, neste caso, o sinal da janela completa sofra duas transições ao longo de sua duração. Isto ajuda a explicar alguns dos fatos observados nas Figuras 26-28, quando da mudança de fase e de período dos sinais. Vale lembrar que este tipo de método é usado de forma muito similar em sistemas comerciais de RTO, como o Romeo, da Invensys (vide Seção 4.1).

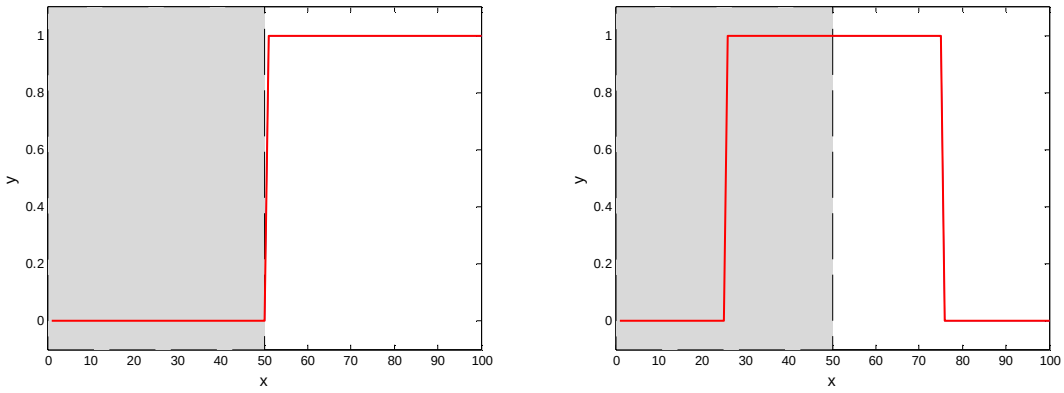


Figura 31 – Comparação do teste de Ackemann-Schladt para dois sinais que sofrem transições súbitas de mesma intensidade. No gráfico da esquerda o método indica não estacionariedade, enquanto que, no da direita, há a indicação de estacionariedade. A área hachurada indica a primeira metade do sinal.

- o método baseado na razão R não pressupõe um formato de variação da forma de onda do sinal nem a comparação de médias entre subjanelas arbitrariamente divididas, mas associa a estacionariedade ao fato de a variância do sinal prover de fontes de natureza aleatória, caso em que seu cálculo independeria da referência usada. Esta verificação é feita por meio da estatística R ou Cs (Equações 224, 226). Como visto nos

casos anteriores, este método mostrou a maior divergência de resultados em relação aos demais. Exceto no caso em que a função teste era uma reta com inclinação variável, nos demais ele indicou persistentemente não estacionariedade. Esta exceção pode ser explicada pela diferente razão sinal/ruído entre os exemplos da função teste reta e periódica. Na verdade, este método é muito sensível a esta razão pois, a depender do período de amostragem do sinal, a variação entre pontos consecutivos pode ser pequena demais face à variabilidade em relação à média, tornando diminuto o valor relativo de s_d^2 e, consequentemente, gerando valores elevados de C. Por exemplo, na Figura 32 mostra-se o efeito da diminuição do intervalo de amostragem para um sinal do tipo y = ax na ausência de ruído aleatório (esquerda) assim como o efeito da relação sinal ruído sobre os valores de C, no caso de contaminação por ruído gaussiano aditivo.

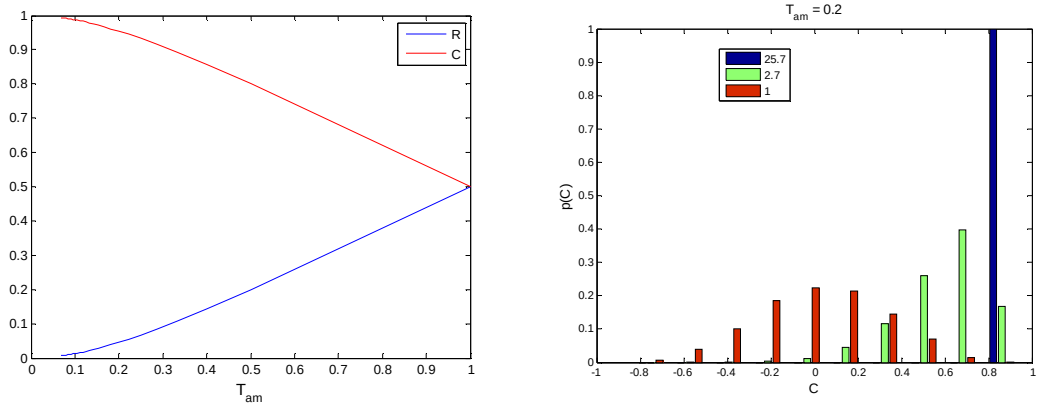


Figura 32 – Esquerda: efeito do período de amostragem no valor de R e C para um sinal do tipo y = ax. Direita: efeito da relação sinal/ruído (cores) sobre a probabilidade de C para y = ax+ε.

Na Figura 33 pode ser visto que, para a função teste senoidal usada no exemplo da Figura 28 (período da função = 1), apenas para valores de ruído bem mais elevados do que os anteriormente utilizados são geradas distribuições de valores de Cs que se superpõem ao valor limite C_{crit} (α=0,05, numero de pontos = 100). Contudo, se suficiente espaçamento é empregado entre amostras consecutivas (Figura 34), s_d^2 torna-se mais importante e Cs se aproxima de zero (assim como R se aproxima de 1).

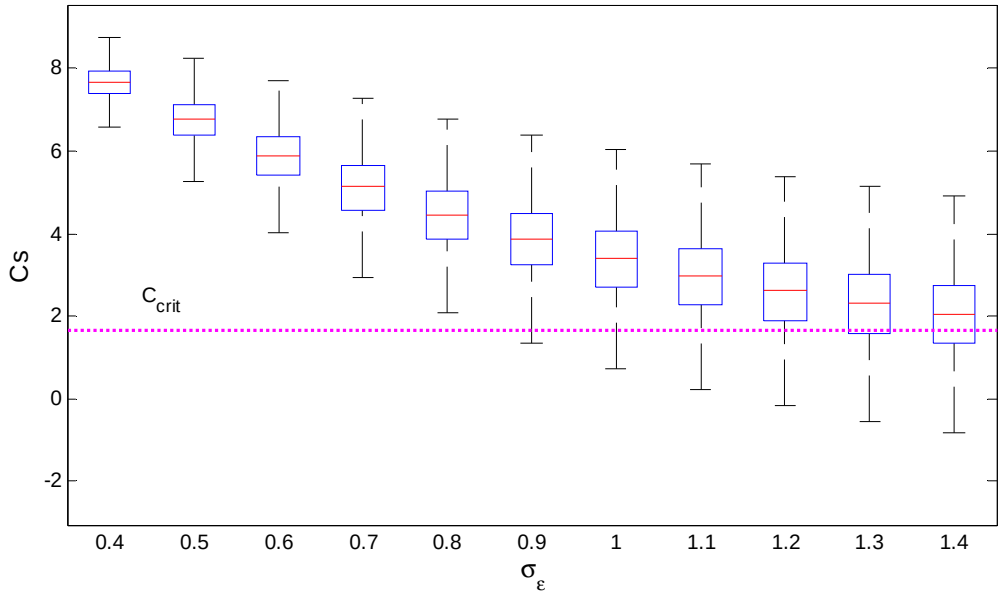


Figura 33 – Comportamento da distribuição da variável Cs para função teste senoidal (amplitude e período unitários, período de amostragem = 0,01, , janela de 100 pontos) contaminado com ruído gaussiano aditivo.

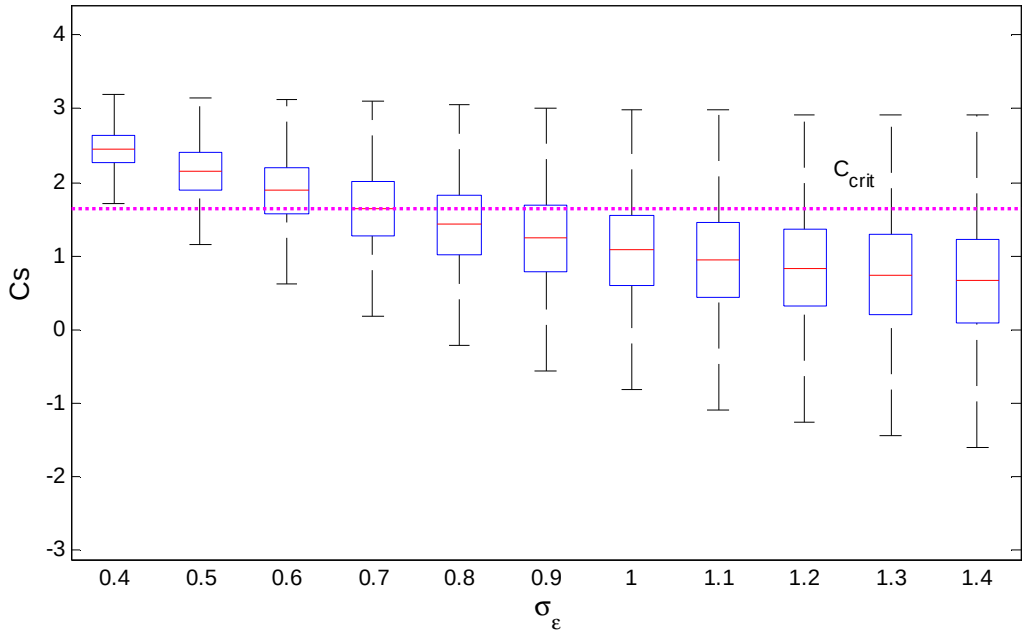


Figura 34 – Comportamento da distribuição da variável Cs para função teste senoidal (amplitude e período unitários, período de amostragem = 0,2, janela de 100 pontos).

A estatística C (ou R) é mais flexível que testes como o do coeficiente de inclinação pois não pressupõe formatos para o sinal. Contudo, ela é muito sensível à relação de dependência entre elementos consecutivos de uma série. Para sinais oriundos de sistemas regidos por equações diferenciais, que produzem informação de natureza contínua e correlacionada, a diminuição progressiva do intervalo de amostragem diminui o valor relativo de s_d^2 em face de s^2, induzindo o aumento do valor de C (225), ou equivalentemente, a diminuição de R (224), levando estes valores para além dos valores críticos e induzindo a indicação de não estacionariedade. Tal fato poderia ser menos importante se s_d^2 apresentasse dependência explícita com análogos discretos da taxa de variação, ao invés de apenas lidar com a variância do sinal.

A sensibilidade ao período de amostragem, bastante pronunciada no teste C, pode ser verificada de forma comparativa com os demais testes estudados através da observação da Figura 35, onde a função teste é a série autogressiva apresentada na Equação (239). Note-se que, à medida que as amostragens tornam-se mais frequentes, a probabilidade de indicação de estacionariedade cai fortemente em relação aos demais métodos em virtude do rápido decréscimo de s_d^2 e da sensibilidade da pdf da estatística C (e de R) a esta queda. A sensibilidade de C também é mais pronunciada em relação ao tamanho da amostra do sinal (tamanho da janela) em relação às estatísticas usadas nos outros métodos, fato que fica evidente da observação da Figura 36. Por estes fatores, apesar da flexibilidade aceita para o sinal de teste, e ainda que seja usado em alguns sistemas comerciais de RTO (AspenPlus versão 5.3), este método apresenta algumas dificuldades de implementação direta na análise de sinais contínuos apropriados por meio de processos de discretização. O pretenso caráter objetivo das conclusões é ofuscado pela obscura interpretabilidade dos resultados em virtude das profundas alterações produzidas por diferentes escolhas de períodos de amostragem.

$$y_{i+1} = a y_i + \varepsilon,\ a = 0{,}35,\ \sigma_\varepsilon = 0{,}25 \qquad (239)$$

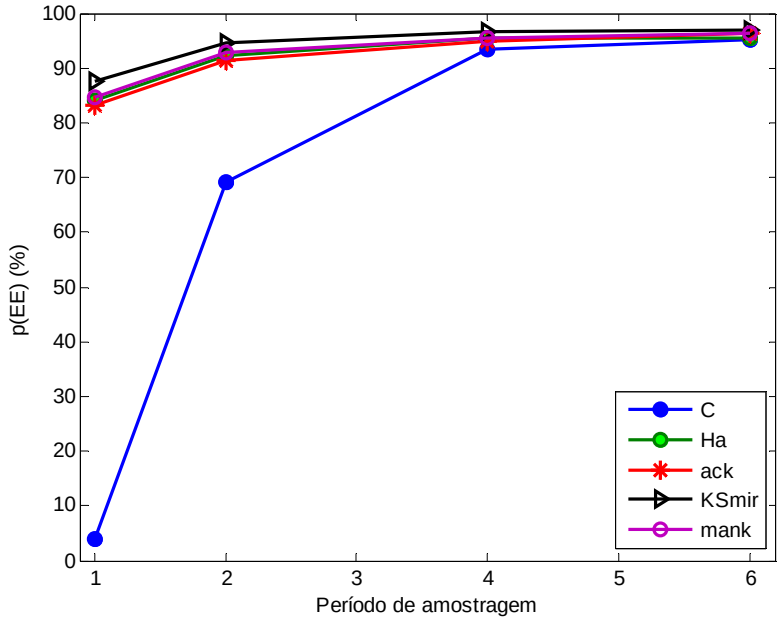


Figura 35 – Influência do intervalo de amostragem (tamanho da janela constante) sobre a probabilidade de indicação de estacionariedade para diversos testes de estacionariedade.

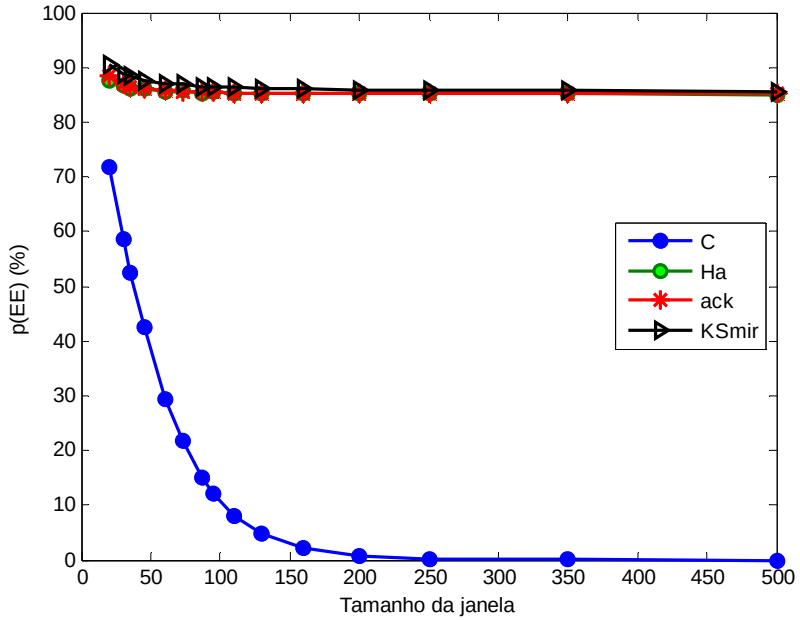


Figura 36 – Influência do tamanho da janela (intervalo de amostragem constante) sobre a probabilidade de indicação de estacionariedade para diversos testes de estacionariedade.

3.2.3.2. Detecção de estacionariedade x Utilidade

Na Seção anterior foi ressaltada a variabilidade dos resultados de diversos testes típicos de estacionariedade em função de variações aparentemente triviais das suas condições de aplicação, assim como a peculiaridade dos resultados apresentados por cada teste.

Na presente Seção será ressaltado o descompromisso entre os resultados dos testes de estacionariedade e os resultados de interesse na execução de um sistema de RTO. Um conceito que deveria ser caro a qualquer projeto de RTO é o da relação de consequência entre as decisões de cada etapa e o "produto" gerado pela execução do RTO. Embora a definição deste "produto", ou a elaboração de métricas que o definam seja objeto de debate, o conceito de que deva haver algum tipo de vínculo entre o teste de estacionariedade e o resultado do RTO deve ser estabelecido desde logo. Este vínculo define o conceito de *utilidade* do teste de estacionariedade. Sem ele, o teste torna-se um conceito vazio, cujo objetivo começa e termina em si mesmo.

Ainda que a existência do conceito de utilidade e o imperativo de seu uso pareçam evidentes, escapa ao registro deste autor a existência de *softwares* comerciais e de aplicações na literatura que dele façam uso ativo. De forma rotineira, os testes são formulados e executados nos moldes apresentados na Seção anterior, sujeitos apenas às suas próprias definições intrínsecas e descompromissados das consequências de sua aplicação. Embora pretensamente se faça uso de ferramentas com justificativa estatística, ou suas premissas estão comprometidas ou os fundamentos dos testes de hipóteses feitos não se relacionam à expectativa do resultado último do sistema de RTO. Como exemplo, pode-se citar a ausência de relação entre a adoção de um determinado nível de significância estatístico e o grau de sub-optimalidade esperado para o valor da função econômica. Em todos os testes comumente usados, tal relação não é evidenciada. Na verdade, o nível de significância é um parâmetro abstrato não correlacionado com o interesse ou a utilidade de seu uso para o serviço prestado pelo RTO.

Desta forma, o conceito genérico de utilidade aqui adotado servirá para responder a seguinte pergunta: Dado um domínio constituído pela classe de sinais esperados para avaliação antes de cada execução do RTO, de que forma a decisão de aceitar um conjunto de sinais como passíveis de serem processados pelo RTO impacta na região esperada de acordo com dado critério de desempenho? Para dar prosseguimento a esta discussão é interessante formalizar determinados conceitos,

alguns dos quais já livremente utilizados anteriormente, como o conceito de janela de dados.

Supõe-se que, em dado instante j, dispõe-se da história pregressa das informações obtidas por observação direta em uma série de n elementos uniformemente separados no tempo por um intervalo T_{am}, de acordo com a amostragem prevista na Equação (48). Neste caso, a série que representa a janela de dados disponível em j é dada por $[\mathbf{Za}]_{j,n,Tam}$ de acordo com a Equação (240), sendo que o intervalo temporal que indica a sua profundidade é dado pela Equação (241).

$$[\mathbf{Za}]_{j,n,Tam} = [\mathbf{Za}_{\text{j-n+1}} \ldots \mathbf{Za}_{\text{j}}] \tag{240}$$

$$\Delta jan_{n,Tam} = (n-1)T_{am} \tag{241}$$

As definições (240,241) explicitam o fato de que a configuração da janela de dados, $[\mathbf{Za}]_{j,n,Tam}$, usada como matéria-prima para o teste de estacionariedade, é mais um item importante no rol de decisões estruturais do RTO. Assim sendo, pode-se atualizar este rol acrescentando ao conteúdo apresentado na Equação (128): a configuração da janela de dados, $J_{,n,Tam}$; o conjunto de índices **ee** (242), que referencia o subconjunto das variáveis obtidas por observação direta que serão usadas para fins de detecção de estacionariedade; e o método de estacionariedade Υ, associado à sua parametrização, $\boldsymbol{\theta}_{EE}$, vinculado à nomenclatura $\Upsilon_{\boldsymbol{\theta}_{EE}}$. O conjunto de escolhas, **Rto**, ficará deste modo caracterizado como na Equação (243).

$$\mathbf{ee} = \{\mathbf{x} \mid \mathbf{x} \subseteq \mathbf{ms}\} \tag{242}$$

$$\mathbf{Rto} = \left\{\mathbf{in}, \mathbf{upd}, \mathbf{obj}, \mathbf{df}, \mathcal{M}_{\boldsymbol{\theta}_e}, J_{n,Tam}, \mathbf{ee}, \Upsilon_{\boldsymbol{\theta}_{EE}}\right\} \tag{243}$$

A janela de dados $[\mathbf{Za}(\mathbf{ee})]_{j,n,Tam}$ contém dim(**ee**) sinais de comprimento n que serão alvo da análise de estacionariedade. A conformação dos sinais desta janela é condicionada pela evolução dos sinais de entrada **Z**(**in**) ao longo do mesmo intervalo de tempo. Ou seja, se for conhecido, a priori, que os sinais contidos na janela de dados

$[\mathbf{Z(in)}]_{j,n,Tam}$ pertencem a determinada classe $\mathscr{C}$ de sinais, formulada com base na expectativa de sua ocorrência no processo físico, pode-se conhecer a família de sinais que contém as informações passíveis de estarem contidas nas janelas $[\mathbf{Za(ee)}]_{j,n,Tam}$.

Se a atuação do RTO for avaliada por uma métrica de desempenho das variáveis de processo e/ou da função objetivo, $\mathbb{M}(\mathbf{ZZ},\mathrm{L})$, a utilidade de um teste de estacionariedade aplicado no domínio configurado pelos sinais de entrada contidos na classe dos sinais $\mathscr{C}$ está associada à expectativa, moldada pelo RTO, do formato da região induzida para $\mathbb{M}(\mathbf{ZZ},\mathrm{L})$. Desta forma, pode-se entender que a utilidade $\mathbb{U}(\mathbb{T}\rightarrow\mathscr{C})$, de um teste de estacionariedade $\mathbb{T}$ em dada classe $\mathscr{C}$ de sinais de entrada do processo seja espelhada pela função distribuição de probabilidade de $\mathbb{M}$ induzida pelos veredictos do teste, conforme a Equação (244). Testes mais úteis são aqueles capazes de produzir, com maior probabilidade, valores de $\mathbb{M}$ mais convenientes à política de condução operacional da planta.

$$[\mathbf{Z(in)}]_{j,n,Tam} \in \mathscr{C} \;\Rightarrow\; \mathbb{U}(\mathbb{T}\rightarrow\mathscr{C}) = \psi\left(\mathbb{M}(\mathbf{ZZ},\mathbf{L})\right) \tag{244}$$

Para exemplificar a importância do conceito de utilidade como discriminador da real função do teste de estacionariedade sob a perspectiva integrada de um sistema de RTO, apresenta-se um caso simples, de mistura em linha de dois produtos enviados a um tanque intermediário, como mostrado na Figura 37. A classe à qual pertencem os sinais é dada por $\mathscr{C}$, e definida de acordo com as Equações (245,246), expressando uma série senoidal restrita a sinais com $T_0<9$, onde T_0 é a variável aleatória associada ao teste de tendência de variação linear H_a, como descrito na Equação (219).

$$\mathscr{C} = \left\{ \left(y = y(0) + d\,\frac{\sum_{i=1}^{N} a_i sen\left(\frac{2\pi t}{b_i} + c_i\right)}{\sum_{i=1}^{N} a_i} \right) \middle| \, T_0(y)<9 \right\} \tag{245}$$

$$0 \le t \le 3000;\; Tam = 15;\; a \sim u(0{,}05;\; 1);\; b \sim u(10;15000);\; c \sim u(0;\pi);\; \mathrm{N} = 5 \tag{246}$$

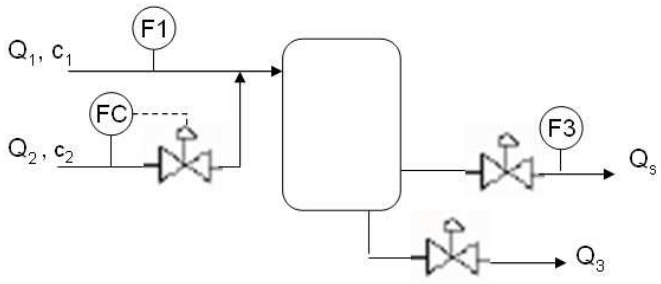


Figura 37 – Diagrama simplificado do processo usado como exemplo para a otimização em presença de não estacionariedade.

O processo é perturbado por variações na vazão da corrente de alimentação mais diluída, $\dot{Q}_1$. As vazões de saída, $\dot{Q}_3$ e $\dot{Q}_s$ são dependentes da diferença de pressão entre os extremos de suas linhas e também da altura do nível do tanque. O processo mostrado na Figura 37 é descrito pelas equações diferenciais, $\boldsymbol{f}_{dif,}$ apresentadas nas Equações (247,248) e pelas equações algébricas $\boldsymbol{f}_{alg}$ (249,250).

Balanço de massa global

$$\dot{Q}_1 + \dot{Q}_2 - \dot{Q}_3 - \dot{Q}_s = A\frac{dh(t)}{dt} \tag{247}$$

Balanço de massa do soluto

$$\dot{Q}_1 c_1 + \dot{Q}_2 c_2 - \dot{Q}_3 c_s - \dot{Q}_s c_s = A\frac{dhc_s}{dt} \tag{248}$$

Equações das válvulas das linhas de saída do vaso

$$\dot{Q}_s = k_s x_v \sqrt{P_{vaso} - P_{descarga1} + \rho g h} \approx K_s \sqrt{h}\,, \tag{249}$$

$$\dot{Q}_3 = k_3 x_{v3} \sqrt{P_{vaso} - P_{descarga2} + \rho g h} \approx K_3 \sqrt{h} \tag{250}$$

O conjunto de variáveis descritivas do processo é representado por **ZZ** em (251). De acordo com a representação proposta, os conjuntos **O e dO** são mostrados, respectivamente, nas Equações (252,253), enquanto $\boldsymbol{\tau}$ é representado pela área da Seção reta do tanque, A (Equação 254). Por outro lado, a representação reduzida ***fm***, o modelo

estático do processo, é descrita pelo sistema de Equações (255) por meio das variáveis descritivas **Zm** (256).

No cenário de operação previsto, as concentrações das correntes de alimentação se mantêm constantes, assim como o parâmetro K_s e a área da seção do tanque, A. Todas as vazões, com exceção de $\dot{Q}_{3,}$ são obtidas por observação direta, via instrumentação. A descrição funcional das variáveis em **Zm** incorporada pelo RTO é mostrada na Equação (257) e resumida na Tabela 5 onde também está mostrado que o modelo é adaptado por meio da modificação do valor da vazão não medida, $\dot{Q}_3$.

A formulação da função objetivo econômica, que é uma função não linear das variáveis medidas $\dot{Q}_1$, $\dot{Q}_2$ e $\dot{Q}_s$, e das variáveis não medidas $\dot{Q}_3$ e c_s é apresentada na Equação (258).

$$\{\mathbf{ZZ}\} = \left\{ \underbrace{\dot{Q}_1 \ \dot{Q}_2 \ K_3 \ K_s \ c_1 c_2 \ \dot{Q}_3 \ \dot{Q}_s \ A}_{\mathbf{II}} \ \underbrace{h \ c_s \ \frac{dh}{dt} \ \frac{d(hc_s)}{dt}}_{\mathbf{OO}} \right\} \tag{251}$$

$$\{\mathbf{O}\} = \{h \ c_s\} \tag{252}$$

$$\{\mathbf{dO}\} = \left\{ \frac{dh(t)}{dt} \quad \frac{dh(t)c_s(t)}{dt} \right\} \tag{253}$$

$$\{\boldsymbol{\tau}\} = \{A\} \tag{254}$$

$$\boldsymbol{fm}: \begin{cases} \dot{Q}_1 + \dot{Q}_2 - \dot{Q}_3 - \dot{Q}_s = 0 \\ \dot{Q}_1 c_1 + \dot{Q}_2 c_2 - \dot{Q}_3 c_s - \dot{Q}_s c_s = 0 \end{cases} \tag{255}$$

$$\{\mathbf{Zm}\} = \left\{ \dot{Q}_1 \ \dot{Q}_2 \ c_1 c_2 \ c_s \ \dot{Q}_3 \ \dot{Q}_s \right\} \tag{256}$$

$$\begin{cases} \mathbf{fix} = [3\,4]^{\mathrm{T}}; \ \mathbf{upd} = [6]; \\ \mathbf{in} = [1 \ 2 \ 3 \ 4 \ 6]^{\mathrm{T}}; \ \mathbf{out} = [5 \ 7]^{\mathrm{T}} \\ \mathbf{ms} = [1 \ 2 \ 7]^{\mathrm{T}}; \ \mathbf{dual} = [7] \\ \mathbf{obj} = [7]; \ \mathbf{df} = [2] \end{cases} \tag{257}$$

Tabela 5 – Classificação das variáveis do RTO

	$\dot{Q}_1$	$\dot{Q}_2$	c_1	c_2	c_s	$\dot{Q}_3$	$\dot{Q}_s$
in	●	●	●	●		●	
out					●		●
ms	●	●					●
ms$^-$	●	●					
dual							●
df		●					
upd						●	

Função objetivo econômica

$$L = \$_s(\dot{Q}_s(t) - \dot{Q}_3(t)) - \$_{\dot{Q}1}\dot{Q}_1(t) - \$_{\dot{Q}_2}\dot{Q}_2(t) \qquad (258)$$

$$\text{se } \dot{Q}_s c_s \le \left(\dot{Q}c_s\right)_0,\ \ \$_s = a_1\dot{Q}_s c_s + b_1$$

$$\text{se } \dot{Q}_s c_s > \left(\dot{Q}c_s\right)_0,\ \ \$_s = a_2\dot{Q}_s c_s + b_2$$

A cada ciclo em que houver a indicação de que o processo se encontra em estado estacionário o RTO deve estimar $\dot{Q}_3$ e sugerir um set-point para a variável de decisão, $\dot{Q}_2$, que otimize a operação do processo sob o ponto de vista da maximização do lucro (258).

Dentre as variáveis necessárias, **Zm**(**in**), a restrição de pertencimento à classe $\mathscr{C}$, de acordo com as Equações (245,246) se aplica à $\dot{Q}_1$, uma vez que $\dot{Q}_2$ é grau de liberdade da otimização.

Serão estudadas as consequências das seguintes decisões estruturais do RTO, concernentes à detecção de estacionariedade:

- **ee**: conjunto de sinais escolhido para a detecção de estacionariedade - será considerado o uso das condições *c1*: **ee**=[1]; *c2*: **ee**=[7]; *c3*: **ee** = [1 7], respectivamente correspondentes às escolhas de ($\dot{Q}_1$); ($\dot{Q}_s$); ($\dot{Q}_1$, $\dot{Q}_s$). Note-se que, para a condição c3, que compreende o uso de mais de um sinal, é usada a regra de combinação que exige a

simultaneidade de detecção de estacionariedade para ambos os sinais de modo a que o teste resultante indique a presença de estacionariedade.

- $\mathcal{T}_{\mathbf{0}_{EE}}$: método de detecção e parametrização – para facilitar a compreensão dos resultados, e aproveitando a similaridade dos resultados gerados pelos diversos testes, serão apresentados preferencialmente os resultados dos testes *Ha* e *ack*. Quando conveniente para a interpretação dos resultados, serão apresentados também os resultados para os testes *mank* e *KSmir*. Sempre que aplicável, será considerado o nível de significância $\alpha = 0{,}05$.

Os resultados apresentados nesta Seção foram obtidos a partir de um conjunto de 5.10^5 sinais de $\dot{Q}_1$ pertencentes à classe $\mathscr{C}$, como definido nas Equações (245,246). A métrica cuja pdf está associada à utilidade dos testes está definida na Equação (259), e consiste no desvio relativo entre o valor da função objetivo econômica atingido com as escolhas efetuadas e o valor ótimo "verdadeiro", L_{otm}, nas condições de operação. Este valor foi calculado baseado na suposição de que as variáveis de entrada mantém-se em valores constantes, no valor médio da janela de dados observados, para todos os instantes para além da implementação da solução proposta pelo RTO.

$$\mathcal{M} : \Delta L(\%) = 100\,\frac{L - L_{otm}}{L_{otm}} \tag{259}$$

As funções distribuição de probabilidade estimadas para ΔL(%), quando usado o teste Ha, podem ser vistas na Figura 38. Pode-se notar que as diferentes escolhas de **ee** produzem diferentes pdf, evidenciando as diferentes *utilidades* associadas a cada escolha. A definição da escolha *mais útil* sempre envolverá certo grau de subjetivade pois, ainda que possa ser dito, de forma qualitativa, que é evidente o interesse em uma distribuição mais concentrada ao redor do valor mais conveniente, ainda assim será necessário quantificar esta concentração, traduzindo a curva contínua da p.d.f. em um parâmetro de discriminação (ex: valor de determinado percentil, distâncias inter-quartis, valor médio etc...).

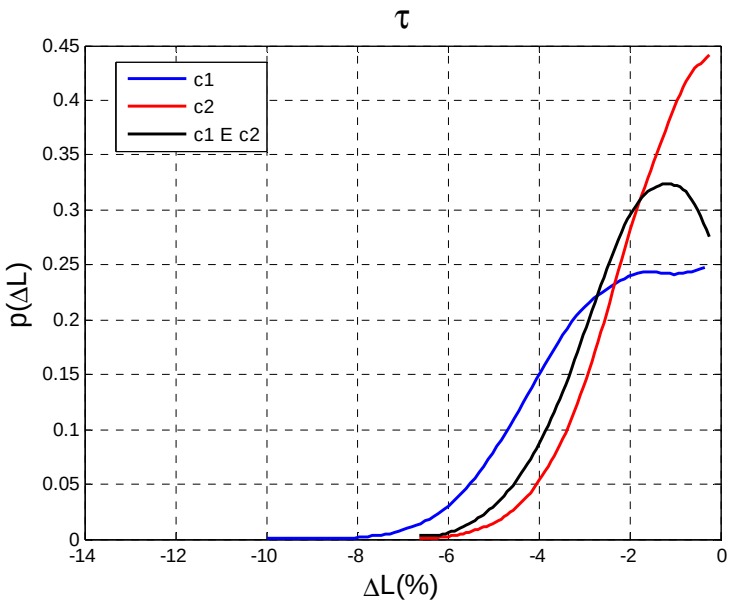


Figura 38 – Função densidade de probabilidade do desvio percentual entre o lucro ótimo da planta e o produzido pelo RTO. Detecção de estado estacionário via verificação de tendência linear, Ha. C1: **ee**=[1]; C2: **ee**=[7];

No caso presente, ainda que de forma qualitativa, a escolha do sinal de $\dot{Q}_s$ como matéria-prima para a execução do teste de estacionariedade aparentemente produz resultados mais convenientes. Contudo, a observação cuidadosa deste fato permite a formulação do seguinte questionamento: como a escolha de uma condição menos restritiva, C2, pode gerar resultados melhores do que o atendimento simultâneo das condições C1 e C2?

A resposta a este questionamento traz à tona uma deficiência de origem de todos os testes de estacionariedade apresentados, qual seja, a do descompromisso entre a formulação do parâmetro de discriminação de estacionariedade (comumente uma variável estocástica a serviço do teste de hipóteses estatístico) e a métrica de interesse último do sistema de RTO. Isto pode ser verificado através da análise dos diversos gráficos contidos na Figura 39, para as condições C1 e C2. Nesta figura pode ser visto que os valores da estatística T para os testes Ha e ack contidos na faixa de validação da estacionariedade não apresentam qualquer correlação com os valores de ΔL(%). Para os casos apresentados, é quase equiprovável que o valor de ΔL(%) associado a determinado percentil seja obtido com valores de T próximos a zero ou próximos aos valores extremos da região de aceitação. Na Figura 40 são apresentados dados similares para a condição C3 (C1 E C2). Nesta figura pode-se observar com mais detalhe os fatos descritos por meio da observação comparativa dos valores brutos dos pares ΔL(%) x T e

de sua apresentação sob a forma de linhas descrevendo os percentis ao longo da faixa de aceitação de T.

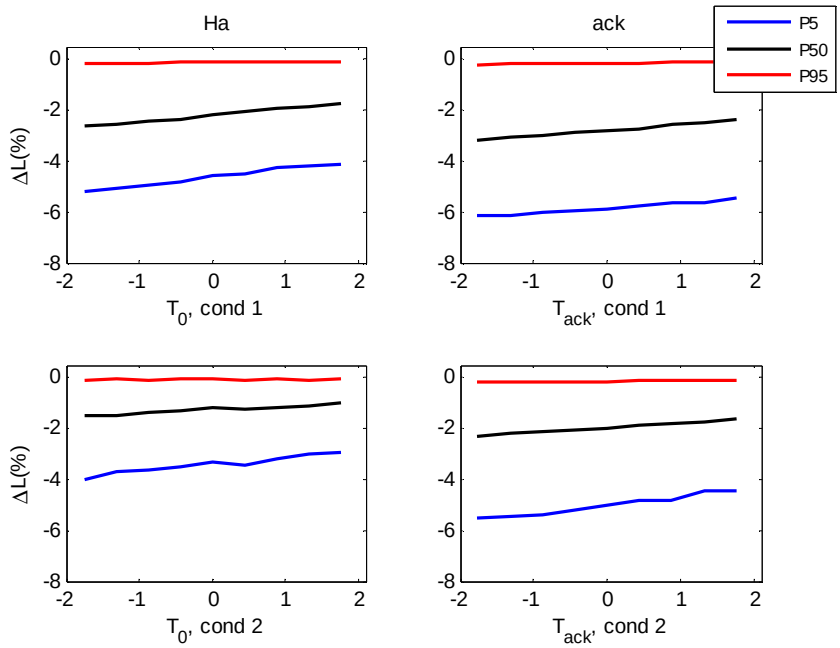


Figura 39 - Relação entre o grau de suboptimalidade e o valor da estatística usada como critério de decisão de estacioneriedade para os testes Ha e ack tomados para as condições 1 e 2. As linhas indicam o 5°, 50° e 95° percentis das distribuições ao longo de 10 segmentos da faixa de valores de T. A=20,Tam=15

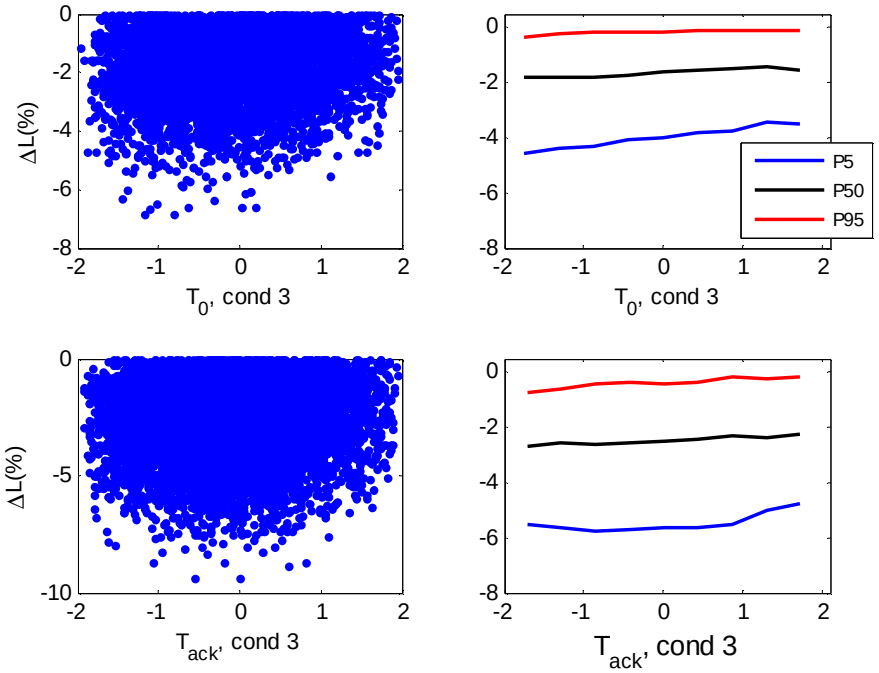


Figura 40 - Relação entre o grau de suboptimalidade e o valor da estatística usada como critério de decisão de estacioneriedade para os testes Ha e ack tomados para o atendimento da condição 3 (c1 E c2). Esquerda: Valores brutos; Direita: As linhas indicam o 5°, 50° e 95° percentis das distribuições ao longo de 10 segmentos da faixa de valores de T. A=20,Tam=15

Voltando à discussão originada com o questionamento proposto, percebe-se que o atendimento simultâneo das condições C1 e C2 tem como único efeito diretamente comprovável *a priori* a diminuição do número de execuções do RTO, na medida em que a probabilidade de que um sinal pertencente à classe $\mathscr{C}$ seja rotulado como estacionário será reduzida em virtude do atendimento combinado de ambas as condições, como será mostrado mais adiante nesta Seção. Contudo, não há nenhuma garantia de que o sinal que atende a C1 e C2 simultaneamente apresente, obrigatoriamente, chances maiores de apresentar valores de $\Delta L(\%)$ mais próximos de zero. Para as condições do caso presente, na verdade, o uso combinado das condições C1 e C2 reuniu o pior dos dois mundos: a diminuição da frequência de execução do otimizador e a indução de $\psi(\Delta L)$ menos útil aos interesses econômicos da planta.

A afirmação feita anteriormente, de que o uso da condição C3 (C1 E C2) torna menos provável a aceitação dos sinais pertencentes a $\mathscr{C}$ e, consequentemente, menos frequente a execução do RTO pode ser vista na Tabela 6 para os testes Ha, ack, mank e KSmir. Contudo, o interesse maior por trás da apresentação destes dados é evidenciar a influência do ruído estocástico que possa estar presente aditivamente ao sinal de $\dot{Q}_1$. Note-se a grande influência do ruído $\boldsymbol{\varepsilon}$ em todos os testes e condições, o que, novamente, diz muito a respeito da susceptibilidade dos testes a circunstâncias alheias à sua real utilidade (e mesmo à possível definição do que seja estacionariedade). Isto se torna ainda mais importante devido ao fato de os sistemas de RTO serem usados em um ambiente industrial, onde esta variabilidade de curto prazo associada a $\boldsymbol{\varepsilon}$ é muito afetada pelos filtros de condicionamento do sinal dos diversos elos da cadeia de medição (sensor, transmissor, SDCD) cuja implementação e parametrização são muitas vezes desconhecidas em sua especificidade pelos usuários finais dos sistemas.

Outra circunstância digna de ser mencionada é a influência da parametrização interna do modelo de processo na utilidade do teste de estacionariedade. Como esta parametrização é invisível ao teste, que foca apenas na morfologia dos sinais, independente do processo ao qual estão conectados, é fácil perceber algumas óbvias consequências deste fato.

Tabela 6 – Probabilidade (%) de indicação de estacionariedade para as três configurações de **ee** (C1,C2,C3), para quatro diferentes testes de estacionariedade e dois diferentes níveis de ruído aleatório aditivo (r0: $\sigma(\boldsymbol{\varepsilon}) = 0$, r2: $\sigma(\boldsymbol{\varepsilon})$ = 2% do valor médio do sinal).

		c1	c2	c3
Ha	r0	21.6	6.7	1.9
	r2	43.9	19.9	12
ack	r0	27.8	9.8	2.8
	r2	54.5	27.2	16
mank	r0	25.6	8.4	2.5
	r2	51.2	22.5	14.8
KSmir	r0	12.5	1.2	0.3
	r2	57.2	30.8	19.8

Suponha-se dois processos distintos, isto é, que transformam de modo diferente a informação contida nas variáveis necessárias medidas em variáveis consequentes, no sentido da Equação (72), mesmo que suas variáveis coincidam integralmente em sua significação física. Se ambos forem estimulados com os mesmos sinais, apesar de a detecção de estacionariedade não se alterar para os estímulos, as consequências produzidas por eles, ao longo das duas camadas consecutivas de otimização do RTO não poderão ser, *a priori*, consideradas idênticas. A Figura 41 mostra, lado a lado, $\psi(\Delta L)$ estimado quando a área (A) do tanque possui o valor nominal e quando este valor é dobrado. O valor de A compõe o vetor $\boldsymbol{\tau}$, como mostrado na Equação (254). Note-se que a maior consequência (que poderia ser facilmente deduzida das premissas dos testes usados) se dá quando é usada a condição c1, que depende inteiramente da morfologia do sinal da variável de entrada, $\dot{Q}_1$. Em ambos os casos as identificações de estacionariedade serão idênticas pois os sinais de $\dot{Q}_1$ são os mesmos. Porém, a propagação pelo RTO é muito distinta. Ainda que neste caso $\boldsymbol{\tau}$ esteja associado a uma característica cuja mudança ao longo da operação não seja fisicamente provável, em um processo genérico a utilidade do método poderá ser modificada a reboque de mudanças em qualquer fator que altere a cadeia de processamento da Equação (72).

A análise do problema exposto na Figura 41 ainda pode ser extendida se considerarmos que $\boldsymbol{\tau}$ contém valores que pertencem a **ZZ**, mas não a **Z.** Ou seja, a Figura 41 apresenta consequencias sobre $\Delta L(\%)$ originadas de variações sofridas por

variáveis que dizem respeito ao problema dinâmico, mas que inexistem na apresentação do problema estacionário. Isto implica que a determinação da utilidade de um método de detecção de estacionariedade requer o conhecimento do modelo dinâmico do processo, ainda que o RTO se atenha apenas ao caso estacionário.

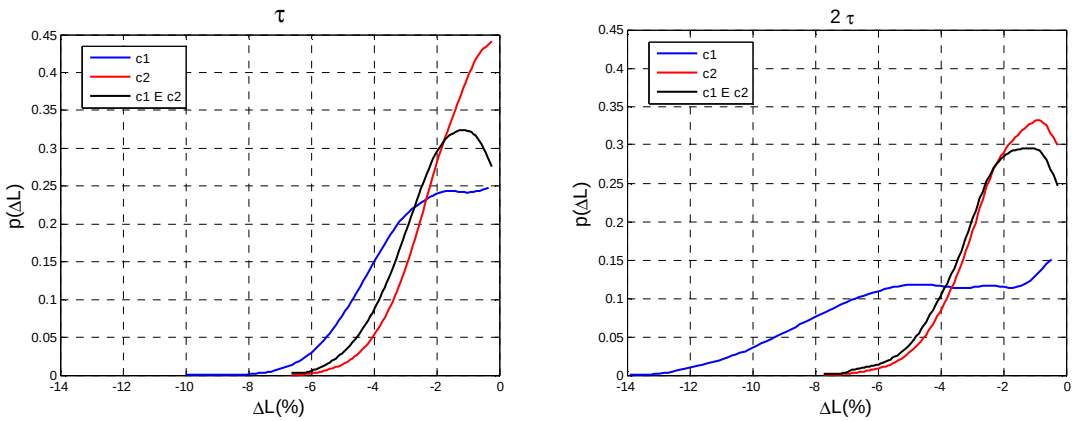


Figura 41 – Função densidade de probabilidade do desvio percentual entre o lucro ótimo da planta e o produzido pelo RTO. Detecção de estado estacionário via verificação de tendência linear, Ha. O gráfico da direita é produzido em um processo cujos valores do vetor $\boldsymbol{\tau}$ são o dobro dos da esquerda.

O próximo ponto a ser explorado é o fato de que alguns aspectos descritivos dos sinais são invisíveis aos testes de detecção de estacionariedade, embora tenham grande impacto na definição da função distribuição de probabilidade das métricas de desempenho. Todos os métodos comumente usados olham para aspectos de natureza relativa à morfologia do sinal, tais como a inclinação do ajuste linear, a variação relativa das médias e desvios-padrões, o pertencimento à mesma pdf ou a variação relativa entre pontos consecutivos. Embora atentos a certos tipos de padrão de formato ou de variabilidade intrínseca dos valores, eles são insensíveis a variações na escala dos sinais, ou seja, pode-se afirmar que sinais de potência distinta produzirão os mesmos veredictos de estacionariedade, ou ainda mais generalizadamente, sinais gerados por transformações lineares de um sinal original compartilharão o mesmo veredicto de estacionariedade para cada teste considerado. Contudo, não há, na formulação do problema de RTO, quaisquer garantias de que a sequência de informações desde $\mathbf{ZZ}_j$ até a poposição de $Lm_{j+1|j}$ (Equação 132), seja a mesma para estímulos que difiram entre si por transformações lineares. Na verdade, em se tratando de processos não lineares, seria surpreendente se assim o fosse.

A afirmação do parágrafo anterior, a respeito da insensibilidade dos testes de estacionariedade à potência do sinal, pode ser mais bem entendida através de alguns exemplos analíticos. Tomando-se como base o teste *ack*, cuja estatística T_{ack} é apresentada na Equação (216), podemos supor que a janela de dados disponível para o teste de estacionariedade consista no vetor $z=[x_1\ x_2\ y_1\ y_2]$, que será dividido simetricamente nas subjanelas $x=[x_1\ x_2]$ e $y=[y_1\ y_2]$, para fins do teste de hipótese de que suas médias sejam iguais, H_0: $\mu_x = \mu_y$.

Colocado o problema desta forma, as estimativas das médias de cada subjanela são calculadas de acordo com a Equação (260), enquanto que as estimativas dos respectivos desvios padrão se apresentam como nas Equações (261, 262). Sob a forma do problema ora proposto, a estatística T_{ack} formulada na Equação (216) apresenta-se como mostrado na Equação (263). Pode ser verificado, acompanhando-se a Equação (264), que transformações lineares da janela de dados **z**, do tipo E**z** + S, resultam no mesmo valor da estatística T_{ack}, comprovando sua insensibilidade a este tipo de transformação e, consequentemente, confirmando que veredictos de estacionariedade idênticos serão assinalados para sinais de potências diferentes.

$$\bar{x} = (x_1 + x_2)/2,\ \ \bar{y} = (y_1 + y_2)/2 \tag{260}$$

$$s_x = \sqrt{\frac{\left(\frac{(x_1 - \bar{x})}{2}\right)^2 + \left(\frac{(x_2 - \bar{x})}{2}\right)^2}{n-1}} = \sqrt{\frac{\left(\frac{(x_1 - x_2)}{2}\right)^2 + \left(\frac{(x_2 - x_1)}{2}\right)^2}{n-1}} = \frac{\sqrt{2}}{2}\sqrt{\frac{(x_1 - x_2)^2}{n-1}} \tag{261}$$

$$s_y = \frac{\sqrt{2}}{2}\sqrt{\frac{(y_1 - y_2)^2}{n-1}} \tag{262}$$

$$T_{ack}([x_1, x_2, y_1, y_2]) = \frac{\sqrt{2}}{2}\frac{x_1 + x_2 - y_1 - y_2}{\sqrt{\frac{(x_1 - x_2)^2}{(n-1)n} + \frac{(y_1 - y_2)^2}{(n-1)n}}} \tag{263}$$

$$T_{ack}\left(E[x_1,x_2,y_1,y_2]+S\right)=\frac{\sqrt{2}}{2}\frac{E(x_1+x_2-y_1-y_2)}{\sqrt{\frac{(Ex_1-Ex_2)^2}{(n-1)n}+\frac{(Ey_1-Ey_2)^2}{(n-1)n}}}=$$
$$=\frac{\sqrt{2}}{2}\frac{x_1+x_2-y_1-y_2}{\sqrt{\frac{(x_1-x_2)^2}{(n-1)n}+\frac{(y_1-y_2)^2}{(n-1)n}}} \quad (264)$$

Complementando esta demonstração, uma análise similar pode também ser feita para o teste de estacionariedade baseado na estatística R (224), partindo-se do sinal z apresentado na Equação (265). O objetivo é saber se uma transformação linear (266) é capaz de alterar o valor da estatística R. Para qualquer sinal fruto desta transformação, o cálculo da média e da variância tradicional é obtido por intermédio das Equações (267, 268), enquanto que a variância calculada a partir de diferenças sucessivas é obtida por meio da Equação (269).

Sendo N=3 o comprimento do sinal z, o valor da estatística R resultará, por meio da combinação de (265-269), na expressão apresentada na Equação (270), onde pode ser claramente percebido que o fator de escala E e o *offset* S introduzidos pela transformação linear (266) são inócuos para o cálculo de R, comprovando sua insensibilidade à potência do sinal.

$$z=[x_1\ x_2\ x_3] \quad (265)$$

$$T(z)=Ez+S \quad (266)$$

$$\bar{z}=\frac{(x_1+x_2+x_3)E+3S}{N} \quad (267)$$

$$s^2=\left(\left(Ex_1+S-\bar{z}\right)^2+\left(Ex_2+S-\bar{z}\right)^2+\left(Ex_3+S-\bar{z}\right)^2\right)\frac{1}{N-1} \quad (268)$$

$$s_d^2=\frac{1}{N-1}\sum_{i=2}^{N}(x_i-x_{i-1})^2=\frac{((Ex_2-Ex_1)^2+(Ex_3-Ex_2)^2)^2}{N-1} \quad (269)$$

$$\mathrm{R} = s_d^2/(2s^2) = \frac{3}{4}\frac{x_1^2 + x_3^2 + 2x_2^2 - 2x_1x_2 - 2x_3x_2}{x_1^2 + x_2^2 + x_3^2 - x_2x_1 - x_1x_3 - x_3x_2} \tag{270}$$

Como visto na Tabela 6, os testes de estacionariedade são sensíveis a padrões de variação relativa, sendo bastante impactados pela razão sinal/ruído, ainda que esta sensibilidade não esteja a serviço da utilidade do teste. Contudo, se a amplitude média dos sinais pertencentes a $\mathscr{C}$ for dobrada, isto não causa nenhuma mudança na probabilidade de que um sinal seja assinalado como estacionário para nenhum dos testes da Tabela 6. Para os exemplos apresentados nesta Seção esta mudança na amplitude média foi obtida mediante a alteração da p.d.f. do parâmetro de amplitude, *d*, na Equação (245). O caso base considera $d = 0{,}2$, enquanto o caso em que a amplitude é dobrada considera $d = 0{,}4$.

Embora a mudança de amplitude média seja invisível aos testes, ou seja, não afete a probabilidade de detecção de estacionariedade, ela certamente afeta a utilidade dos testes, como pode ser visto pela alteração da função distribuição de probabilidade de ΔL(%), como mostrado na Figura 42. O panorama pode ser expandido para os quatro testes de estacionariedade abordados nesta Seção mediante análise da Figura 43, que repete o mesmo padrão de alteração na utilidade e constância da identificação.

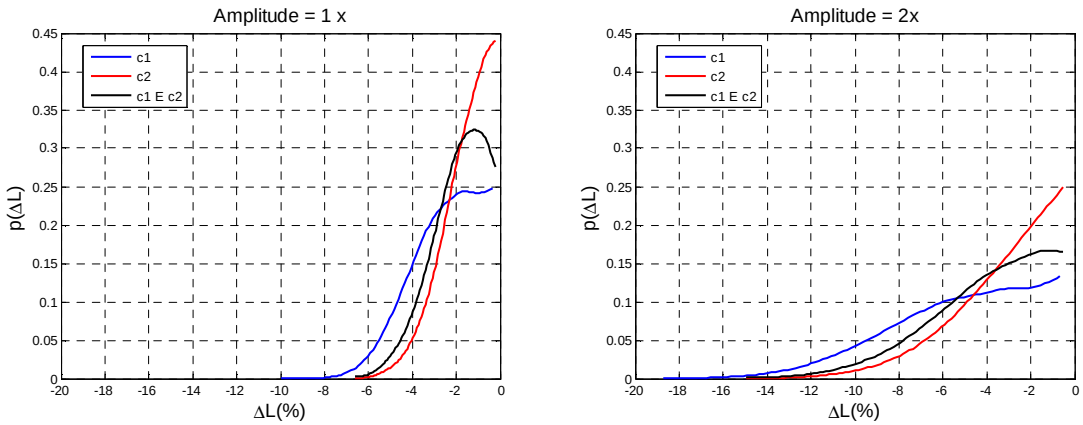


Figura 42 – Função densidade de probabilidade do desvio percentual entre o lucro ótimo da planta e o sugerido pelo RTO para dois diferentes valores de amplitude média de $\dot{Q}_1$. Detecção de estado estacionário via verificação de tendência linear, H_a. C1: **ee**=[1]; C2: **ee**=[7];

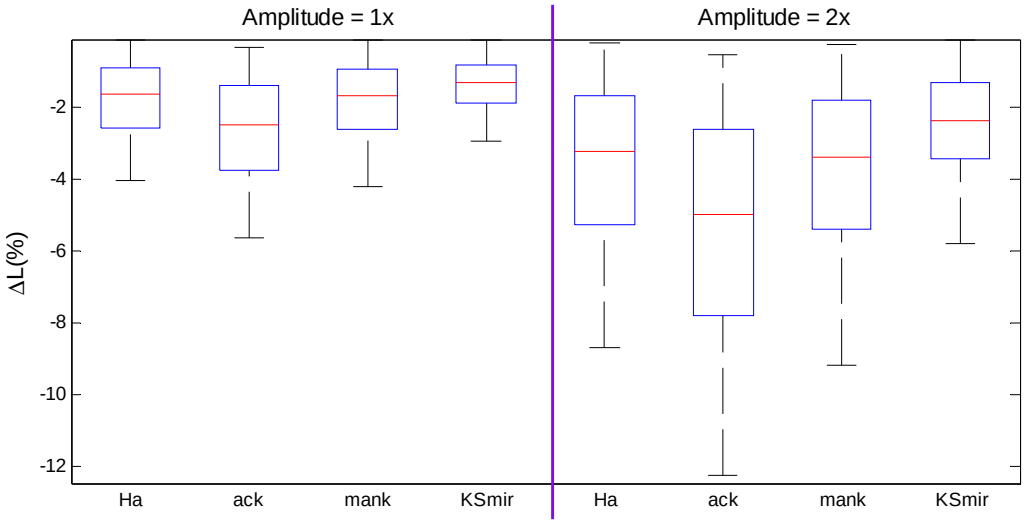


Figura 43 – Distribuição dos valores de ΔL(%) para dois diferentes valores de amplitude média de $\dot{Q}_1$. São apresentados os resultados esperados para quatro diferentes métodos de detecção de estacionariedade.

Em síntese, apresentou-se, nesta Seção a fundamentação que deveria nortear o uso e a escolha das decisões estruturais pertinentes à detecção da estacionariedade, compondo o conjunto completo de decisões apresentado na Equação (243). Esta fundamentação passa necessariamente por uma medida de utilidade que vincule, de forma consequente, as decisões feitas aos resultados que justificam a existência do sistema de RTO. A Tabela 7 resume alguns dos aspectos discutidos, referentes à influência de parâmetros do modelo dinâmico e à insensibilidade dos testes de estacionariedade à potência dos sinais.

Tabela 7 – Valor do percentil 5% de probabilidade de ΔL(%) para os três casos de escolha de **ee** descritos no texto, C1, C2, C3 = (C1 E C2) para variações do modelo do processo e da amplitude média do sinal de entrada. Método H_a.

	C1	**C2**	**C1 E C2**
(τ, Ampl 1x)	-4.7	-3.5	-4.0
(τ, Ampl 2x)	-10.4	-7.4	-8.7
(2τ, Ampl 1x)	-9.4	-3.9	-4.3

3.2.4. Alternativas para o controle da execução do RTO

Claro está que o problema da caracterização da estacionariedade passa pela definição de parâmetros que não se atenham apenas à morfologia do sinal, mas que correlacionem bem os níveis de tolerância assumidos com o grau de sub-otimalidade esperado para as previsões da função objetivo econômica. Além disto, devem ser previstas regras de combinação dos parâmetros de cada sinal de modo a caracterizar o processo globalmente, uma vez que ficou mostrado, a partir da discussão contida na Seção 3.2.3, que a típica regra de combinar os resultados individuais do teste de cada um dos sinais por meio de operações lógicas não é útil, no sentido anteriormente definido. Outro ponto importante é a escolha, voltada para a utilidade, de qual subconjunto dos sinais, **ee**, deve ser usado para os testes.

Este tipo de abordagem do problema de estacionariedade, focado estritamente nas consequências, não é comum na literatura, exceção feita ao trabalhos de Poulin, Hodouin e Lachance [98], no âmbito de reconciliação de dados, e de Flehmig e Marquardt [99], no contexto da otimização. Este último é o único trabalho encontrado na literatura que faz uma crítica similar em alguns aspectos à contida nesta tese, a respeito da ausência de conexão entre os conceitos embutidos nos testes de estacionariedade e a consequência para a otimização econômica final. No caso de Flehmig e Marquardt [99], visa-se determinar uma razão de proporcionalidade entre a violação de estacionariedade observada nas variáveis medidas em relação a não estacionariedade oculta nas variáveis não medidas do processo. Deste modo, seria possível inferir o desvio da estacionariedade não diretamente expresso na morfologia dos sinais oriundos das medições, auxiliando na tarefa de fixar a tolerância para os desvios observados. Contudo, a abordagem de Flehmig e Marquardt [99] requer o uso de um modelo linearizado e não supõe a presença das restrições de desigualdade, ***gm*** (Equação 13), no problema do RTO, o que implica certa idealização das condições de implementação.

A definição de um parâmetro de estacionariedade e de um patamar de tolerância que sejam úteis ao RTO consiste em um aspecto importante da proposta de Flehmig e indicam um caminho promissor. Sua conclusão ratifica o que foi proposto na Seção anterior: que é necessário dispor de um modelo dinâmico do processo, ainda que o RTO refira-se ao problema de otimização em estado estacionário.

O modo de definir e detectar a estacionariedade, associado à respectiva tolerância e regra de decisão multivariável pode ser responsável por uma parcela considerável da

qualidade global das soluções propostas pelo RTO. Neste contexto, uma etapa fundamental é a de definir o modo como se usa a informação pregressa contida na janela $[\ ]_{j,N,Tam}$, destinada à análise de estacionariedade e à adaptação do modelo. O modo mais apropriado estatisticamente consiste em realizar-se a adaptação via maximização da função de verossimilhança sujeita às restrições impostas pelo modelo completo (dinâmico) do processo, como mostrado na Equação (271). Contudo, o modo de atuação do RTO estático faz uso da representação reduzida (Equação 104) que contempla apenas a descrição em termos estacionários e que faz o processo de adaptação ser representado como na Equação (272).

$$\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\max}\left(F\left(\underbrace{[\mathbf{ZZm(obj)}]_{j,N,Tam}}_{\text{mod}elo}, \underbrace{[\mathbf{ZZa(obj)}]_{j,N,Tam}}_{observação} \right) \right)$$

$$s.a \qquad (271)$$

$$\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \end{cases}$$

$$\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\max}\left(F\left(\underbrace{[\mathbf{Zm(obj)}]_{j,N,Tam}}_{\mathbf{Z}_{\text{mod}elo}}, \underbrace{[\mathbf{Za(obj)}]_{j,N,Tam}}_{\mathbf{Z}_{obs}} \right) \right)$$

$$s.a \qquad (272)$$

$$\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \\ \mathbf{dO} = \mathbf{0} \\ \boldsymbol{\tau} = \mathbf{0} \end{cases}$$

De modo a ilustrar a consequência desta representação reduzida sobre os resultados da atuação do RTO, será usado como exemplo o processo descrito de forma completa pelas variáveis **ZZ** (273), inter-relacionadas pela equação diferencial ordinária de 1ª ordem (274). O análogo estático de representação deste processo faz uso das variáveis do modelo **Zm** (275) sujeitas à equação algébrica (276). A instrumentação e a configuração do processo de adaptação do RTO são dimensionadas conforme apresentado em (277) e o RTO deve manipular *x* de modo a minimizar a função objetivo L (278).

$$\{\mathbf{ZZ}\} = \left\{ x \;\; p \;\; A \;\; y \;\; \frac{dy}{dt} \right\} \tag{273}$$

$$f: \quad A\frac{dy}{dt} = px - y \tag{274}$$

$$\{\mathbf{Zm}\} = \{x \;\; p \;\; y\} \tag{275}$$

$$fm: \quad 0 = px - y \tag{276}$$

ms = [1 3], **upd =** [2], **dual** = [3], **obj** = [3] (277)

$$L = c_1 y^2 + c_2 x\,, \qquad \mathbf{df} = [1] \tag{278}$$

A informação contida em **Zm** no instante j é representada pela Equação (279), incorporando informações da observação direta para *x*, do processo de adaptação para *p* e do modelo de processo (276) para a variável dual, *y*. Em contraponto a $\mathbf{Zm}_j$, o seu análogo *verdadeiro*, $\mathbf{Z}_j$ (280), contém as informações reais (assinaladas por um asterisco), onde o valor de y_j provém da Equação *f* (274).

$$\mathbf{Zm}_j = \left[x_j \quad \underbrace{p_0 + \overbrace{\Theta_j(2)}^{\theta}}_{p_j} \quad \left(p_j . x_j\right) \right]^T \tag{279}$$

$$\mathbf{Z}_j = \left[x_j^* \;\; p^* \;\; \underbrace{\left(p^* x_j^* - A\frac{dy}{dt}\bigg|_j \right)}_{y_j^*} \right]^T \tag{280}$$

Se o processo de observação direta produzir **Za** por meio da contaminação **ε** oriunda de uma distribuição normal (281) com matriz de covariância diagonal, o processo de adaptação pode ser representado pela Equação (282), onde a função objetivo

F resume-se ao somatório dos quadrados das diferenças entre **Za** e **Zm** para as variáveis **obj** ao longo da janela de dados. No presente caso, supondo apenas a corrupção da observação de y (283), o vetor de adaptação $\boldsymbol{\theta} = \boldsymbol{\Theta}(2)$ é originado do processo de otimização expresso na Equação (284).

$$\mathbf{Za}_j(\mathbf{ms}) = \mathbf{Z}_j(\mathbf{ms}) + \boldsymbol{\varepsilon} \tag{281}$$

$$\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\min}\left(\sum_{k=j-N+1}^{j} \sum_{i=1}^{\dim(\mathbf{obj})} \left(\mathbf{Za}_k(\mathbf{obj}(i)) - \mathbf{Zm}_k(\mathbf{obj}(i))\right)^2 \right)$$

$$s.a \tag{282}$$

$$\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \\ \mathbf{dO} = \mathbf{0} \\ \boldsymbol{\tau} = \mathbf{0} \end{cases}$$

$$\boldsymbol{\varepsilon}([1\ 2]) = \mathbf{0};\ \boldsymbol{\varepsilon}(3) \sim N(0, \sigma_y) \tag{283}$$

$$\theta = \underset{\theta}{\arg\min} \underbrace{\sum_{k=j-N+1}^{j} \left(\left(p^* x_k - A \left.\frac{dy}{dt}\right|_k \right) + \varepsilon_k - (p_0 + \theta) x_k \right)^2}_{F} \tag{284}$$

Analiticamente, o ponto extremo de *F* que soluciona a Equação (284) corresponde ao valor de θ que torna nula a derivada expressa na Equação (285). Este valor é calculado de acordo com a Equação (286). A correspondência deste valor com um ponto de mínimo de F é validada pela garantia que a segunda derivada de F apresente valor positivo, o que é evidente pela inspeção da Equação (287).

$$\frac{dF}{d\theta} = \sum_{k=j-N+1}^{j} -2x_k \left(\left(p^* x_k - A \left.\frac{dy}{dt}\right|_k \right) + \varepsilon_k - (p_0 + \theta) x_k \right) \tag{285}$$

$$\theta = \Theta(2) = p^* - p_0 - \frac{\mathrm{A}\sum_{k=j-N+1}^{j} x_k \left.\frac{dy}{dt}\right|_k}{\sum_{k=j-N+1}^{j} x_k^2} + \frac{\sum_{k=j-N+1}^{j} x_k \varepsilon_k}{\sum_{k=j-N+1}^{j} x_k^2} \tag{286}$$

$$\frac{d^2F}{d\theta^2} = \sum_{k=j-N+1}^{j} x_k^2 \tag{287}$$

Para fins de clareza na apresentação das relações de causa e efeito, será estudado preliminarmente o caso em que as medições são perfeitas, o que corresponde ao parâmetro de adaptação calculado de acordo com a Equação (288). Note-se que, ainda que o sinal seja observado sem corrupção de sua informação, o modificador θ adaptará de forma errônea o processo em virtude do termo E_{dy}, cuja existência é devida à não estacionariedade do processo.

$$\theta = \Theta(2) = p^* - p_0 - \underbrace{\frac{\mathrm{A}\sum_{k=j-N+1}^{j} x_k \left.\frac{dy}{dt}\right|_k}{\sum_{k=j-N+1}^{j} x_k^2}}_{E_{dy}} \tag{288}$$

O núcleo da influência da não estacionariedade sobre o desempenho do RTO está, portanto, contido no termo E_{dy}, que responderá pelo desvio na estimativa do parâmetro p, conforme a Equação (289). O problema da camada superior de otimização, $x \to \min L$, que está vinculado à função objetivo L (278), será diretamente afetado pelo valor do parâmetro p, como visto por meio das soluções analíticas do problema de otimização (290,291). Este desvio se propaga pela segunda camada de otimização (292), fazendo com que o RTO conduza o processo para um patamar associado ao grau de sub-optimalidade ΔL (294), atingido pela manipulação do grau de liberdade por valor desviado de Δu (293) em relação a seu valor ideal.

$$\Delta p_j = \underbrace{p_j}_{p_0+\Theta(2)} - p^* = \underbrace{-\frac{\mathrm{A}\sum_{k=j-N+1}^{j} x_k \left.\frac{dy}{dt}\right|_k}{\sum_{k=j-N+1}^{j} x_k^2}}_{E_{dy}} \tag{289}$$

$$x_{otm} = -\frac{1}{2}\frac{c_2}{c_1 p^2} \tag{290}$$

$$L_{otm} = -\frac{1}{4}\frac{c_2^2}{c_1 p^2} \tag{291}$$

$$\left[\frac{dy}{dt}\right]_{j,N,Tam} \neq \mathbf{0} \Rightarrow (E_{dy} \neq 0) \Rightarrow (\Delta p \neq 0) \Rightarrow (\Delta u \neq 0) \Rightarrow (\Delta L \neq 0) \tag{292}$$

$$\Delta u = x_{j+1}\Big|_j - x_{j+1}^* = \frac{1}{2}\frac{c_2}{c_1}\left(\frac{1}{p^{*2}} - \frac{1}{\left(p^* - E_{dy}\right)^2}\right) \tag{293}$$

$$\Delta L = L_{j+1}\Big|_j - L_{j+1}^* = \frac{1}{4}\frac{c_2^2}{c_1}\left(\frac{1}{p^{*2}} - \frac{1}{\left(p^* - E_{dy}\right)^2}\right) \tag{294}$$

As expressões analíticas (289, 293, 294) explicitam e quantificam a dependência do grau de sub-optimalidade, ΔL, e do desvio dos graus de liberdade, Δu, com a presença de derivada não nula de y ao longo da janela de dados. Note-se que este fato comprova as observações feitas na Seção 3.2.3.2, sobre a dependência da utilidade com:

- a potência dos sinais, evidente pela dependência de E_{dy} com Σx_k^2,

- a configuração específica do problema em cada instante, evidente pela dependência de Δp, Δu e ΔL com o valor do parâmetro p no instante j e com os parâmetros c_1 e c_2 da função objetivo, L.

- o conhecimento do modelo dinâmico do processo, evidenciado pela presença de elementos do vetor $\boldsymbol{\tau}$ (A na formulação de E_{dy}).

Como já discutido, tais dependências são ignoradas por métodos baseados na morfologia do sinal e não atentos ao modelo interno do problema, como aqueles descritos na Seção 3.2.2. Será apresentado, nesta Seção, um exemplo de método alternativo aos métodos comumente usados para detecção de estacionariedade, como forma de superar as deficiências apresentadas. Em termos generalizados, pode-se pensar nas seguintes características que tal método deva possuir:

- seus veredictos devem dizer respeito às consequências da aceitação dos sinais disponíveis sobre a métrica de desempenho do sistema.

- a parametrização deve ter significado diretamente vinculado à métrica de desempenho, de modo que sua escolha possa ser guiada por sua consequência real.

- deve prover a associação multivariável dos sinais de modo conveniente à utilidade do resultado

Na verdade, em virtude do pragmatismo do uso do método, expresso pela permanente conexão com sua utilidade, é de pouca importância denominá-lo como um teste de dectecção de estacionariedade. Ao invés disto, ele será preferencialemente definido como um *teste de adequação* dos sinais ao uso no RTO. O conceito de estacionariedade, em si, é irrelevante para as finalidades propostas, e surgirá, se presente, como consequência de sua adequação.

Conceitualmente, o método de detecção de adequabilidade de uma janela de dados ao RTO será proposto baseado na Figura 44, e que resume os seguintes procedimentos: 1) descrição dos sinais da janela de dados por meio de características pertinentes; 2) síntese de um veredicto de adequabilidade baseado nas características de todos os sinais disponíveis. Em termos gerais, esta Figura também poderia ser usada para representar muitos dos métodos anteriormente apresentados. Contudo, para estes métodos, cada sinal usualmente é resumido a apenas uma característica (inclinação, média da janela etc..) e a síntese é feita de modo individual, sendo o veredito final dado por uma combinação dos veredictos individuais por meio de um operador *booleano*.

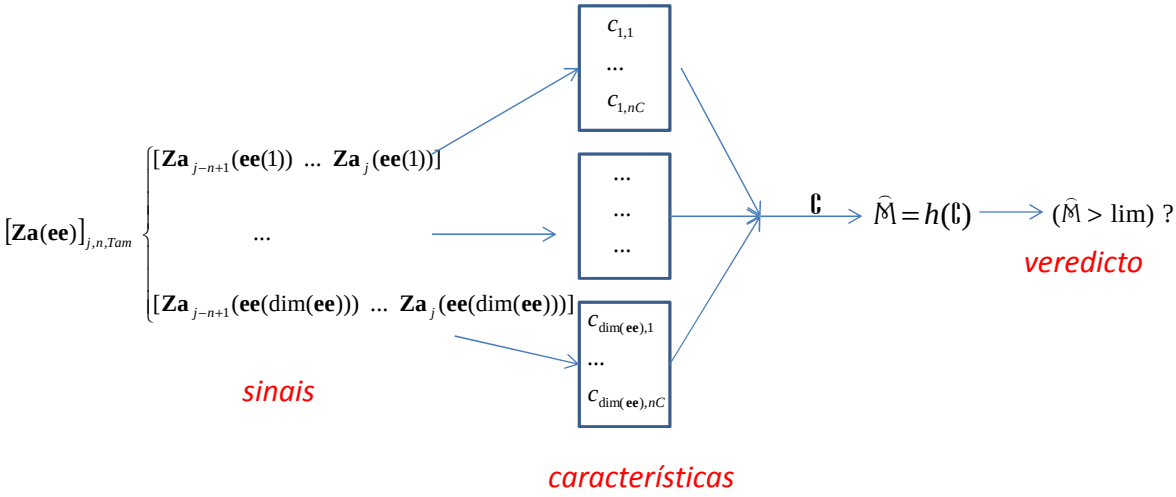


Figura 44 – Esquema conceitual para o método de detecção de adequabilidade.

A extração de características representa um processo de redução da dimensionalidade da análise, a partir do qual os (dim(**ee**) x n) valores contidos na janela $[\mathbf{Za}(\mathbf{ee})]_{j,n,Tam}$ são *trocados* por (dim(**ee**) x nC) valores, supostamente equivalentes em termos de representação para fins de análise de adequabilidade, onde nC representa o número de características analisadas. Esta troca se dá com o objetivo de simplificar a análise, embora, no mais das vezes, esta redução se dê de forma extrema, com nC = 1 (uma característica global para toda a janela) para os testes convencionalmente usados. É fácil verificar que o uso de apenas uma característica global pode ser contraproducente para sistemas de RTO regidos, em sua essência, por equações algébrico-diferenciais (18-20), pois diversas formas de onda podem ser traduzidas por uma mesma estatística global, embora possam produzir resultados muito distintos se estimularem o mesmo processo. É de se esperar, portanto, que a menos que nC>1, qualquer método proposto fique prisioneiro de possíveis ambiguidades na descrição dos sinais.

Isto posto, a primeira questão a ser colocada diz respeito à natureza das características que descreverão o conteúdo da janela de dados. Dada a natureza serial das informações e o fato que as variáveis fazem parte de um sistema de equações diferenciais, é conveniente que, no rol de características descritoras, sejam incluídas informações sobre a localização temporal dos traços da morfologia dos sinais que, de alguma forma, e em alguma medida, sejam necessárias para a representação acurada da utilidade.

Na ausência de uma teoria geral e definitiva sobre a redução da dimensionalidade da janela de dados, a abordagem empregada neste trabalho para lidar com este problema possui natureza exploratória, guiada pelo conhecimento do inventário de técnicas de processamento de sinais e de aspectos estatísticos descritivos como forma de escolha e seleção das características representativas da informação disponível.

Ao contrário dos métodos anteriormente descritos, não se advoga, *a priori*, o conhecimento das propriedades morfológicas condicionantes da adequação do sinal ao seu uso. Ao contrário, este conhecimento é obtido *a posteriori*, baseado em análise da utilidade e fundamentado na premissa de que as características descritoras têm importância relativa para cada problema e para cada classe de sinais, dada em função da especificidade e da configuração das relações de igualdade e desigualdade que descrevem o processo (11-13).

De acordo com esta premissa, as características apresentadas na Figura 44 são escolhidas a partir do conjunto $\mathbf{C}$ (295), que contém o elenco de todas as propriedades, *pp*, candidatas a serem descritoras do conteúdo da janela de dados e, assim, fazerem parte do conjunto de caracterítiscas dos sinais, $\mathbb{C}$. A busca por representações esparsas e compactas é uma preocupação recorrente na literatura técnica de processamento de sinais. No presente caso, o conjunto $\mathbf{C}$ pode ser visto de forma similar a um dicionário de átomos de representação [100,101], embora com a distinta diferença que o objetivo não é encontrar uma base para a reconstrução do sinal, mas sim para compor a representação equivalente de uma métrica de utilidade do RTO que é impactada pelos sinais.

A i-ésima propriedade contida em $\mathbf{C}$, $pp_{i,m}$, é aplicada de forma segmentada, ao longo de *m* subjanelas consecutivas da janela de dados original (296), sendo que, quando *m*=1, a propriedade é calculada de forma global para toda a janela de dados. No limite oposto, se *m*=*n*, cada subjanela contém apenas um ponto amostrado. As restrições contidas em (295) indicam que m deve ser um submúltiplo do número de pontos da janela de dados ($n=\alpha m$) e que cada subjanela deve conter mais de um valor amostrado ($\alpha>1$).

$$\mathbf{C}=\left\{x \mid x=pp_{i,m},\ i=1..npp,\ n=\alpha m,\ (\alpha,m,n)\in\mathbb{N}^{*},\ \alpha>1\right\} \tag{295}$$

$$pp_{i,m}\Big|_{[\,]_{j,n}} = \left\{ \begin{array}{l} x \,|\, x = pp_i([a \;\; b]), \\ a = j + kn/m \\ b = j - 1 + (k+1)n/m \\ k = 0...(m-1),\; k \in \mathbb{N} \end{array} \right\} \tag{296}$$

Como apresentado na Figura 44, a detecção da adequabilidade dos sinais contidos na janela de dados para uso pelo RTO baseia-se na representação reduzida dos sinais por meio das características $\mathfrak{C}$ e pela sua tradução, por meio da relação funcional h($\mathfrak{C}$), em $\hat{\mathcal{M}}$, aproximação da métrica de desempenho do RTO, $\mathcal{M}$. Deste modo, a formulação do método de detecção da adequabilidade pressupõe:

- a formulação preliminar do elenco de potenciais propriedades descritivas, $\mathbf{C}$

- a seleção das características, $\mathfrak{C} \subset \mathbf{C}$, que indicam as propriedades e o respectivo nível de segmentação que formam a representação reduzida dos sinais contidos na janela de dados de análise

- a formulação da relação funcional, h($\bullet$), apresentada na Equação (297), que prediz o valor da métrica de desempenho em função das características descritivas de todos os sinais da janela de dados.

$$\begin{array}{l} h : \mathbb{R}^{\dim(\mathbf{ee}) \,\mathrm{x}\, \dim(nC)} \rightarrow \mathbb{R} \\ \mathfrak{C} \mapsto \hat{\mathcal{M}} \end{array} \tag{297}$$

A definição do conjunto de propriedades candidatas, $\mathbf{C}$, é o ponto de partida do processo, sendo amparada no conhecimento prévio da importância e da especificade das propriedades descritoras. A formulação de $\mathfrak{C}$ e h pode ser expressa como o resultado do problema de otimização apresentado na Equação (298), cuja solução define:

- o vetor **i,** que assinala as propriedades selecionadas em $\mathbf{C}$,

- o vetor de segmentação **m**,

- a relação funcional *h* pertencente a um domínio de funções definido a priori.

A função objetivo do problema de otimização descrito em (298) contém, além da métrica de distância entre a medida de desempenho e sua predição, um termo de parcimônia, dado por γ, cuja função é incorporar uma penalidade à função objetivo que crie uma relação de compromisso entre a complexidade da representação da janela e a qualidade da aproximação da métrica de desempenho.

$$
\begin{aligned}
&(\mathbf{i},\mathbf{m},h) \;=\; \underset{\mathbf{i},\mathbf{m},h(\bullet)}{\arg\min}\left(\left\|\mathcal{M} - h(\mathcal{C})\right\|\gamma\right) \\
&s.a \\
&\begin{cases}
\mathcal{C} \;= \mathbf{C}\big|_{i=\mathbf{i},\; m=\mathbf{m}} \\
h \in \{h_1 \;\ldots\; h_{nh}\} \\
\gamma = \gamma\left(\dim(\mathcal{C}), \max(\mathbf{m}), \mathcal{M}, h(\mathcal{C})\right) \\
\dim(\mathcal{C}) = \dim(\mathbf{i}) \le npp \\
\left[\mathbf{Z}(\mathbf{in})\right]_{j,n,Tam} \in \mathcal{C}
\end{cases}
\end{aligned}
\qquad (298)
$$

Como exemplo comparativo, o teste de adequabilidade organizado segundo a Figura 44 e operacionalizado de acordo com a Equação (298) será aplicado ao problema de RTO descrito pelas Equações (247-258) e sujeito à classe de sinais, $\mathcal{C}$, apresentada nas Equações (245,246), problema este que já foi objeto de estudo na Seção 3.2.3.2 sob a perspectiva dos métodos tradicionais de detecção de estacionariedade. Para este estudo, o *pool* de características descritoras, dado pelo conjunto $\mathbf{C}$ em (298), é formado pelo elenco de propriedades *pp* ao longo dos segmentos das subjanelas. Estas propriedades pertencem às seguintes categorias:

- momentos estatísticos (média, desvio-padrão)
- coeficientes de aproximação polinomial
- descrição temporal do espectro de potência do sinal

Em relação a esta última categoria vale a pena uma discussão adicional a respeito dos requisitos que devem ser atendidos por estas características. A inclusão deste tipo de descrição se dá em virtude da constatação de que a natureza serial das informações é importante dada a estrutura matemática do processo ser descrita por um sistema de

equações diferenciais, em que é pressuposta a dependência dos estados com as condições pregressas do sistema. A localização temporal dos eventos ao longo da janela de dados é de relevante significado pois, para este tipo de sistema, estímulos como os mostrados na Figura 31 causam consequências distintas, apesar de poderem produzir características globais similares.

Um requisito do elenco de características escolhidas para descrever os sinais é de que ele reduza a dimensão da representação original. Contudo a escolha, para o caso geral, de uma base que represente de forma compacta a informação original não é um problema que possa ser resolvido de forma universal e *a priori*, mas que deve ser adequado às especificidades de cada problema. A busca por representações esparsas e compactas de sinais é uma preocupação recorrente em processamento de sinais e uma escolha usual para a análise e processamento de sinais é dada pelo uso de uma base ortonormal de funções trigonométricas, obtido pela expansão em uma série de Fourier que origina, quando devidamente representada sob a forma exponencial, a transformada descrita pelo produto interno dado em (299).

$$X(\omega i) = \left\langle x(t), e^{-i\omega t} \right\rangle \tag{299}$$

A série dada pelos termos $\left|X(i\omega)\right|^2$ constitui o espectro de potência do sinal e informa a distribuição da energia do sinal ao longo de suas frequências componentes. Uma representação que queira levar em conta esta distribuição poderia incorporar os termos desta série como propriedades contidas em $\mathbf{C}$. O valor do espectro para dada janela de dados, tomado para cada ω, representa informação sobre a contribuição de uma frequência específica para a energia total do sinal. Um ponto a ser ressaltado, contudo, é que a elevada precisão da informação da frequência dos coeficientes da série de Fourier ($\Delta\omega \rightarrow 0$) corresponde à completa imprecisão em termos de localização temporal ($\Delta t \rightarrow \infty$), de acordo com o princípio de incerteza enunciado por Gabor [102], o qual prevê que o produto $\Delta t \Delta\omega$ tenha valor constante. Conforme anteriormente mencionado, a falta de localização temporal pode ser um inconveniente para o presente uso.

Uma alternativa comumente usada [103] para contornar este tipo de problema consiste em se empregar uma base de representação de suporte finito para a transformação, ou seja, realizar a projeção do sinal indicada pelo produto interno em (299) em uma base que possua valor não nulo apenas sobre um intervalo finito (janela).

À medida que este intervalo de suporte diminui, aumenta a localização temporal da informação produzida, embora diminua a localização da frequência. A transformada resultante através da aplicação de uma janela de largura ℓ_g , descrita por uma função *g* (300), centrada no instante t_c, é mostrada na Equação (301).

$$\begin{cases} g(t_c + \delta) > 0, \ |\delta| \leq \ell_g / 2 \\ g(t_c + \delta) = 0, \ |\delta| > \ell_g / 2 \end{cases} \tag{300}$$

$$X(\omega i, t_c) = \left\langle x(t), e^{-i\omega t} g(t - t_c) \right\rangle \tag{301}$$

A transformação apresentada na Equação (301), convencionalmente chamada de *Short Time Fourier Transform* (STFT), é um avanço em termos de caracterização, embora ainda sofra de deficiências originadas do fato de que, dada a largura constante da janela, a imprecisão da localização no tempo e na frequência também se mantém constantes para todo o domínio de representação. Isto faz com faz com que não se possa adaptar a resolução para diferentes regiões de análise. Uma vez escolhida a largura e o formato da janela, não há como simultaneamente conciliar uma boa resolução tanto em altas como em baixas frequências. Uma das duas faixas será definitivamente condenada a uma análise deficiente.

Uma evolução natural da idéia por detrás da STFT é considerar a análise multi-escala do sinal, conseguida por meio da adaptação da largura da janela para propiciar, com resolução compatível, a análise em diversos níveis de frequência. Isto é conseguido se a projeção do sinal for feita sobre uma base de suporte finito e ajustável em função da região de interesse da análise. Para atingir tal objetivo, vamos supor inicialmente uma função Ψ que possua as seguintes propriedades [104]: energia finita e média nula, de acordo com a Equação (302).

$$\begin{cases} \int_{-\infty}^{\infty} |\Psi(t)|^2 \, dt < \infty \\ \int_{-\infty}^{\infty} \Psi(t) \, dt = 0 \end{cases} \tag{302}$$

A função Ψ(t) que atende aos requisitos acima, doravante chamada ‘wavelet-mãe’, é o núcleo a partir do qual serão criadas variações, as ‘wavelet-filhas’, como

mostradas nas Equações (303) e (304), respectivamente associadas a funções contínuas e discretas. Estas variações da função principal são ortonormais entre si e construídas por intermédio de duas operações simples: de variação de escala, ou de *dilatação*, por alterações do parâmetro a; de deslocamento temporal, ou de *translação*, por meio de mudanças no parâmetro b. A transformada de wavelet (305) consitirá nos coeficientes da projeção do sinal sobre versões da wavelet-mãe em múltiplas escalas e transladadas ao longo da duração do sinal. O asterísco assinala o complexo conjugado, realçando o fato de que a wavelet-mãe pode ser uma função real ou complexa.

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \tag{303}$$

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{2^a}}\Psi\left(\frac{t-b2^a}{2^a}\right) \tag{304}$$

$$X_W(a,b) = Xw_{a,b} = \left\langle x(t),\ \Psi^*_{a,b}(t)\right\rangle \tag{305}$$

O uso de uma faixa contínua de valores de *a* e *b* implica em elevado esforço computacional assim como gera uma quantidade de dados muito grande e com alto grau de redundância. A discretização dos valores de dilatação e translação é um modo mais conveniente de lidar com sinais reais. A representação contida na Equação (304) incorpora o uso de um grid diádico [104, 105], que nada mais é do que a discretização dos valores da dilatação com base em potências de dois. Conhecidos os coeficientes X_W, o sinal pode ser reconstruído de acordo com a Equação (306).

$$x(t) = \sum_a \sum_b \underbrace{\left\langle x(t),\ \Psi_{a,b}(t)\right\rangle}_{Xw_{a,b}} \Psi_{a,b}(t), \qquad a,b \in \mathbf{Z} \tag{306}$$

Contudo, cada vez que a escala é aumentada em um fator de dois na escala diádica, a largura temporal do suporte da wavelet é dobrada e a largura de frequência é reduzida pela metade, sendo fácil perceber que, à medida que este processo avança, um número cada vez maior de wavelets-filhas são necessárias para preencher espaços cada vez menores no espectro de baixas frequências. Este tipo de problema apresentado pelas

representações discretas de escalas é resolvido [105] pelo uso de uma função cujo espectro cubra o setor inicial [0 f_0] de frequências (sendo f_0 uma frequência arbitrária). Esta função, ϕ é chamada de função de escala (ou wavelet-pai) e o espectro por ela coberto, [0 f_0], corresponde às escalas [a_0 a_1 ... a_∞], sendo a_0 a escala arbitrária escolhida para sua representação. Deste modo é possível [104] representar um sinal genérico em termos de uma representação em dada escala a_0 expressa pela composição de um sinal de aproximação (A) e um sinal de detalhamento (D), como mostrado na Equação (307).

$$x(t) = \underbrace{\sum_{b=-\infty}^{\infty} \left\langle x(t),\, \phi_{a_0,b}(t) \right\rangle \phi_{a_0,b}(t)}_{A} + \underbrace{\sum_{a=-\infty}^{a_0} \sum_{b=-\infty}^{\infty} \left\langle x(t),\, \Psi_{a,b}(t) \right\rangle \Psi_{a,b}(t)}_{D} \qquad (307)$$

A representação do sinal por meio de (307) é a chave para a análise em multiresolução, permitindo a decomposição do sinal em vários níveis de detalhamento e precisão. Esta análise será a base a partir da qual serão compostas as características do sinal, no escopo da análise de adequabilidade do sinal ao RTO. A Figura 45 ajuda a esclarecer as consequências da representação (307). Como pode ser observado, a decomposição faz o sinal passar por um filtro passa-baixa (função de escala), que gera o sinal de aproximação cuja energia está concentrada na região inferior do espectro de potencia, [0 f_0], assim como por um filtro passa-faixa (função wavelet), que gera o sinal de detalhamento que representa os componentes de frequências superiores do espectro de potencia do sinal original. Cada sinal de aproximação pode ser sucessivamente decomposto pelo mesmo processo, gerando representações de detalhamento em diferentes escalas, de modo que $A_{n-1} = A_n + D_n$. No limite, o aprofundamento da divisão em escalas poderia continuar até o extremo em que cada janela fosse reduzida a um ponto, embora na prática sejam levadas em conta as características conhecidas do sinal ou empregado algum critério de entropia para definir o número máximo de níveis de detalhamento. Ao final do processo de decomposição, o sinal é representado por uma aproximação e diversos detalhamentos, de modo que $S = A_n + \sum_{n=1}^{n} D_n$.

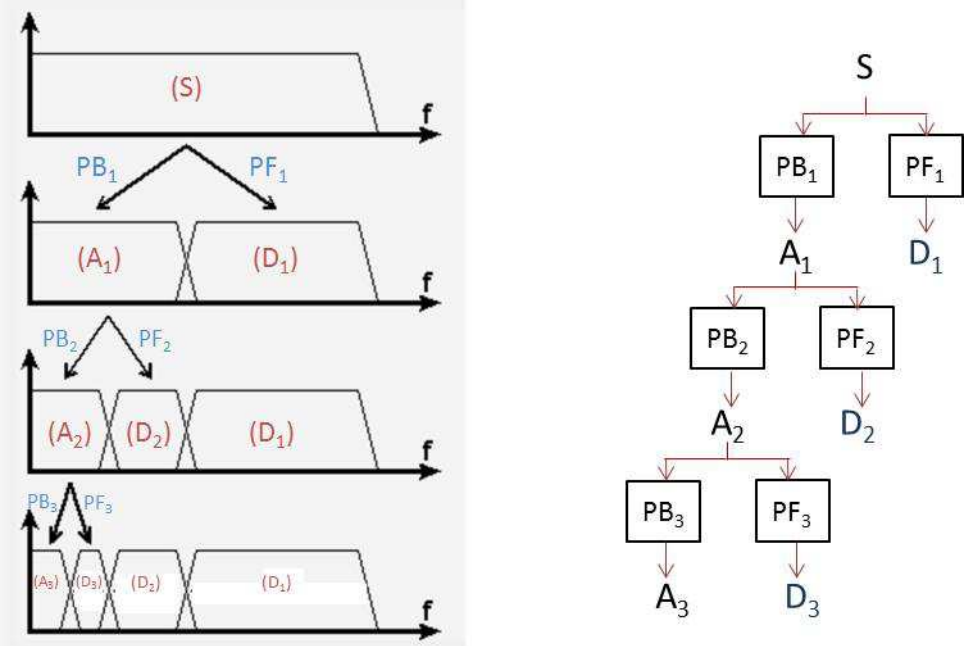


Figura 45 – Decomposição de um sinal em aproximações e detalhes. À direita: banco de filtros de funções de wavelet (passa-faixa, PF) e de escala (passa-baixa, PB). À esquerda: espectro de potência dos sinais produzidos sucessivamente em cada escala de representação.

O processo de decomposição sucessiva pode ser mais bem observado por meio de um exemplo simples, onde a técnica é aplicada a um sinal composto por três frequencias distintas com diferentes localizações temporais, como apresentado na Figura 46. O componente de baixa frequência está presente na primeira metade de sua duração enquanto que o componente de mais elevada frequência está presente na metade restante. Um terceiro componente, de frequência mediana, está presente no terço intermediário, com superposições em ambas as regiões extremas. A decomposição em uma escala diádica discreta em quatro níveis é apresentada na Figura 47. Se observarmos o sinal sob a perspectiva da representação até o quarto nível de escala, o sinal original pode ser equivalentemente descrito por $A_4+D_4+D_3+D_2+D_1$. Note-se que, se procurarmos as características do sinal original, notar-se-á que o componente de baixa frequência está retido no sinal de aproximação, o componente de frequência intermediária aparece primordialmente nos detalhes de 4º e 3º níveis e o componente de frequência superior está retido nos detalhes de 2º e 1º níveis.

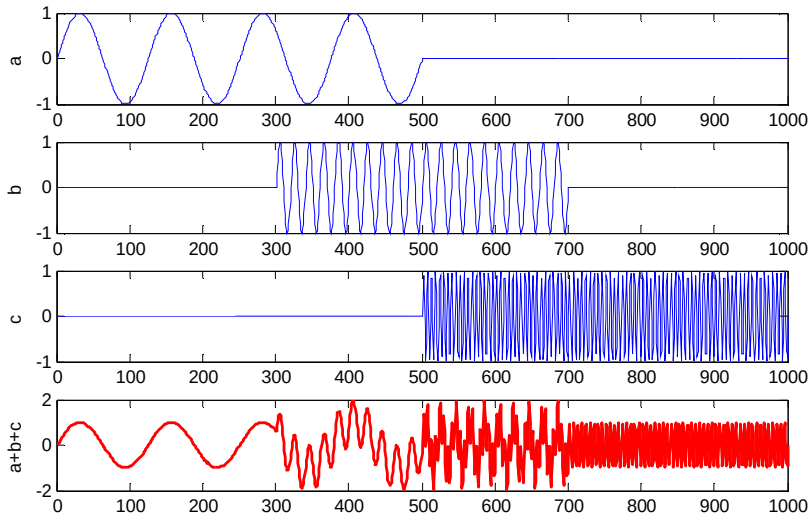


Figura 46 – Sinal de exemplo (vermelho) composto pela adição dos sinais a, b e c.

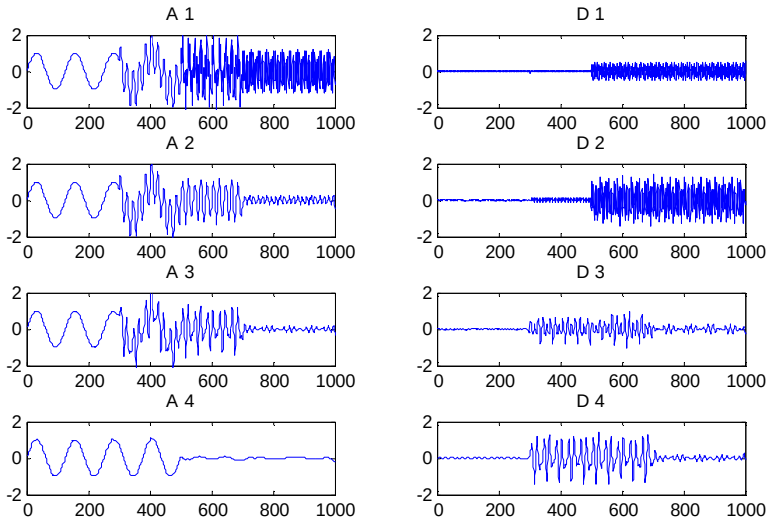


Figura 47 – Representação do sinal de exemplo em termo de sucessivas aproximações e detalhes em uma escala diádica.

Após este interlúdio, onde foram apresentadas as justificativas referentes à escolha de propriedades derivadas do processamento dos sinais por meio de wavelets, podemos retornar ao problema de exemplificar o método de testar a adequabilidade dos sinais contidos em uma janela de dados multivariável ao uso em sistemas de RTO. Como dito anteriormente, pretende-se testar a adequabilidade de sinais pertencentes à classe $\mathscr{C}$ (245,246) ao RTO implementado no processo da Seção 3.2.3.2. Para este problema, o conjunto de características potenciais, $\mathbf{C}$, foi escolhido de modo a conter

propriedades relacionadas a seis níveis de detalhamento e à aproximação de sexto nível de escala do sinal original, obtidas a partir de wavelets da família Daubechies [101] de índice 4. Os sete sinais de análise (seis de detalhamento e um de aproximação) contidos na janela de dados de cada uma das duas variáveis medidas ($\dot{Q}_1, \dot{Q}_s$) são segmentados em tantas subjanelas quanto o especificado no vetor $\mathbf{m} = [m_{\dot{Q}_1}\ m_{\dot{Q}_s}]$. São usadas propriedades estatísticas dos sinais obtidos pela decomposição por wavelets em diversas formas e combinações. No total, para lidar com os sinais de $\dot{Q}_1$ e $\dot{Q}_s$ são criadas cerca de sessenta propriedades potenciais para comporem o conjunto $\mathbf{c}$.

As características descritoras ℂ, calculadas para cada subjanela de cada sinal são incorporadas pela função de aproximação da métrica de desempenho de modo a possibilitar a discriminação da adequabilidade do sinal. Todas as especificações (dimensionamento do número de subjanelas para cada sinal, seleção das características descritoras e da função de aproximação da métrica de desempenho) necessárias à consecução dos procedimentos indicados na Figura 44 são obtidos por meio da solução do problema de otimização descrito em (298). A função que aproxima a métrica de desempenho, *h*, neste problema consiste em uma rede neural feed-forward com dim(ℂ) neurônios na camada de entrada e 8 neurônios em sua camada oculta.

Após a definição da estrutura e do método, a qualidade da detecção de adequabilidade dos sinais ao uso no RTO pode ser avaliada na Figura 48 em termos da métrica de desempenho em (245,246), que representa o desvio percentual, ΔL(%), entre o valor da função objetivo econômica proposta pelo RTO e seu melhor valor possível, conforme apresentado na Equação (259). Nela estão representadas a distribuição do afastamento entre $\mathbb{M} = \Delta L(\%)$ e $\hat{\mathbb{M}} = dL(\%) = h(\mathbb{C})$, que é o valor da métrica de desempenho prevista pelo método de detecção de adequabilidade com base nas características ℂ dos sinais apresentados ao processo e pertencentes à classe $\mathscr{C}$. O desempenho é apresentado de forma separada para quatro casos diferenciados em termos do parâmetro de amplitude d em (245), que está associado à potência do sinal, e do desvio-padrão percentual do ruído gaussiano presente no sinal medido. Desta forma, a simbologia empregada na Figura indica que um sinal que pertença ao caso d*X*s*Y* pertence à classe definida em (245,246) com o parâmetro d=X e ruído gaussiano aditivo igual a Y% do valor médio do sinal. Pelas distribuições apresentadas nota-se que as previsões estão concentradas em torno de ±0,5% de desvio da métrica real.

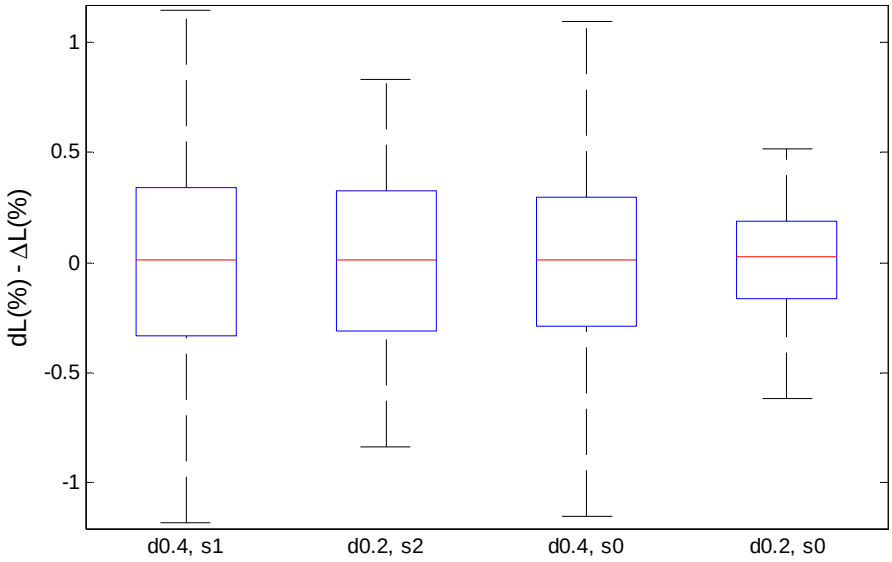


Figura 48 – Diferença entre o afastamento percentual do lucro ótimo previsto pelo método de adequabilidade (dL) e o afastamento real (ΔL), obtido caso um sinal seja aceito para ser processado pelo RTO. Apresentadas quatro combinações de parâmetro de amplitude (d) e desvio-padrão do ruído relativo (s), vide texto.

Um problema que acometia os métodos convencionais, e que era um reflexo de sua falta de compromisso com a utilidade era a ausência de correlação entre seus indicadores e o desempenho final do RTO. Como o método de adequabilidade foca exatamente na previsão do impacto sobre o desempenho, é de se esperar que tal fato não ocorra no presente caso, o que se comprova ao observar-se a Figura 49, onde está mostrada a relação entre o desempenho realizado e aquele predito. É interessante notar o profundo contraste entre a sensibilidade dos valores de ΔL(%) ao resultado do teste de adequabilidade, mostrada na Figura 49, e o comportamento apresentado nas Figuras 39 e 40 na página 130, que apresentam informação análoga produzida pelos testes de detecção de estacionariedade convencionais.

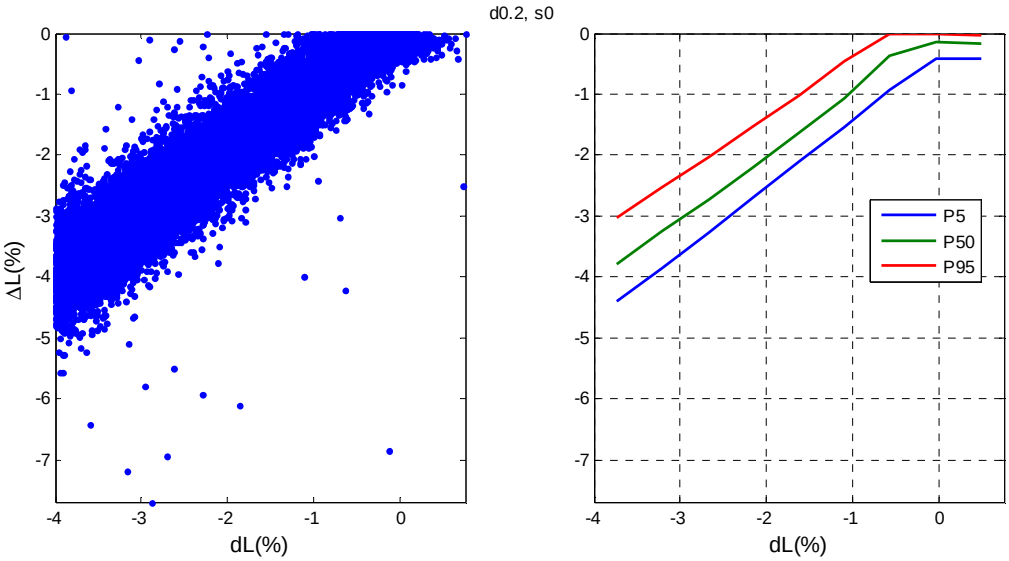


Figura 49 – Correspondência entre o valor da métrica de desempenho calculado pelo método de adequabilidade, dL(%), e o valor real, ΔL(%), para sinais da classe analisada. À esquerda são vistos os valores individuais de cada uma das 5.10^4 simulações. À direita são mostradas as linhas que delimitam o 5°, 50° e 95° percentis dos valores do gráfico à esquerda.

Sob o ponto de vista do emprego real do método de adequabilidade, é necessário caracterizá-lo em termos de seu potencial de falhas, dadas as escolhas feitas na sua configuração e uso. Uma vez que as decisões de adequabilidade serão tomadas baseadas em um critério de desempenho cujo interesse está em impedir a ultrapassagem de valores extremos, a preocupação recai em que sejam julgados adequados sinais que façam o RTO produzir um desempenho na direção oposta. Por exemplo, se forem selecionados sinais com base no critério inferido dL>x, a probabilidade de ocorrência de resultados em que o RTO produza o resultado real ΔL<x é uma informação relevante no uso prático, pois indica a chance de que a decisão tomada não produza o resultado de desempenho esperado. Para o caso presente, isto é materializado no exemplo apresentado na Figura 50. Nela pode ser vista a função de distribuição de probabilidade de ΔL condicionada à seleção de sinais em que $dL > -4\%$. A probabilidade de que a seleção de adequabilidade produza resultados contrários ao interesse econômico do processo corresponde à área hachurada sob a curva, que corresponde a casos em que $\Delta L < -4\%$, a despeito da seleção de sinais baseados na premissa oposta. O desempenho do método pode ser melhor verificado a partir da análise da Tabela 8, onde são apresentados os valores da probabilidade de que valores de ΔL menores que os usados na parametrização do método sejam obtidos para sinais pertencentes a $\mathscr{C}(245,246)$. Pode

ser visto que a probabilidade de desencontro nas previsões aumenta à medida que a seleção dos sinais é feita almejando a faixa superior de desempenho prometido (dL mais próximo de zero).

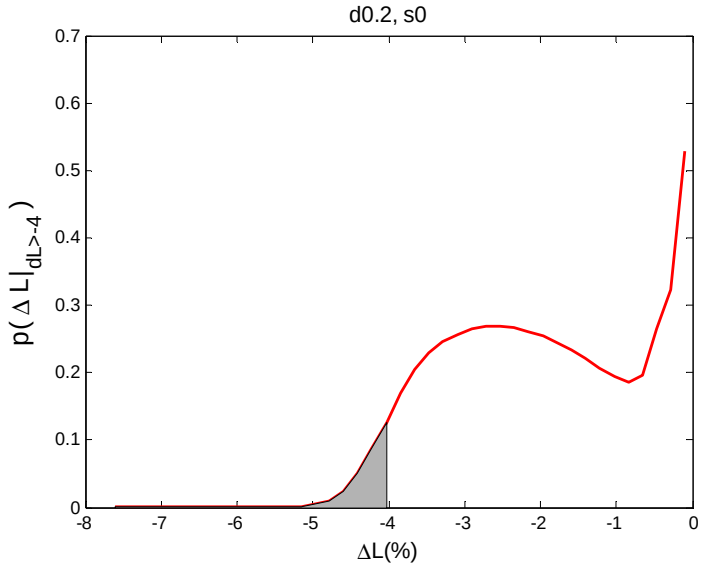


Figura 50 – Função distribuição de probabilidade de ΔL(%) condicionado à ocorrência de sinais em que dL>-4%

Tabela 8 – Probabilidade de ocorrência de desempenho do RTO pior do que o suposto pelo uso do método de adequabilidade. Seleção dos sinais feita com base em cinco diferentes valores de dL. São apresentados os resultados globais da classe de sinais considerada no texto assim como resultados para cada uma quatro combinações de amplitude (d) e de ruído gaussiano (s) isoladamente.

	$p(\Delta L \mid_{dL>x} < x)$				
x(%)	**global**	d0.2, s0	d0.2, s2	d0.4, s0	d0.4, s1
-5	**0.05**	0.03	0.04	0.06	0.06
-4	**0.07**	0.06	0.07	0.07	0.06
-3	**0.10**	0.08	0.10	0.08	0.07
-2	**0.10**	0.09	0.12	0.07	0.08
-1	**0.13**	0.11	0.14	0.09	0.09

Esta discussão pode ser abordada sob um ponto de vista mais amplo a partir da análise da Figura 51. O caso nela apresentado considera que as janelas de dados são selecionadas sob o critério de adequabilidade dL>x, ou seja, dos quais se espera que o desvio relativo do lucro ótimo não seja menor que x, sendo x um número não positivo, o

que decorre da definição do desvio relativo (259). Dados estes sinais selecionados, é mostrada a probabilidade de que apresentem desvio real, ΔL, para além do valor y, ou seja, $p(\Delta L|_{dL>x} < y)$. Se o valor x for um parâmetro do método à escolha do usuário, o dilema a ser enfrentado fica aparente nesta figura ao verificar-se que, embora a escolha de valores mais restritivos (x→0) para a seleção dos sinais aptos ao RTO torne mais garantida a presença de pequenos desvios, a contrapartida é a drástica redução do número de vezes em que o RTO será permitido operar.

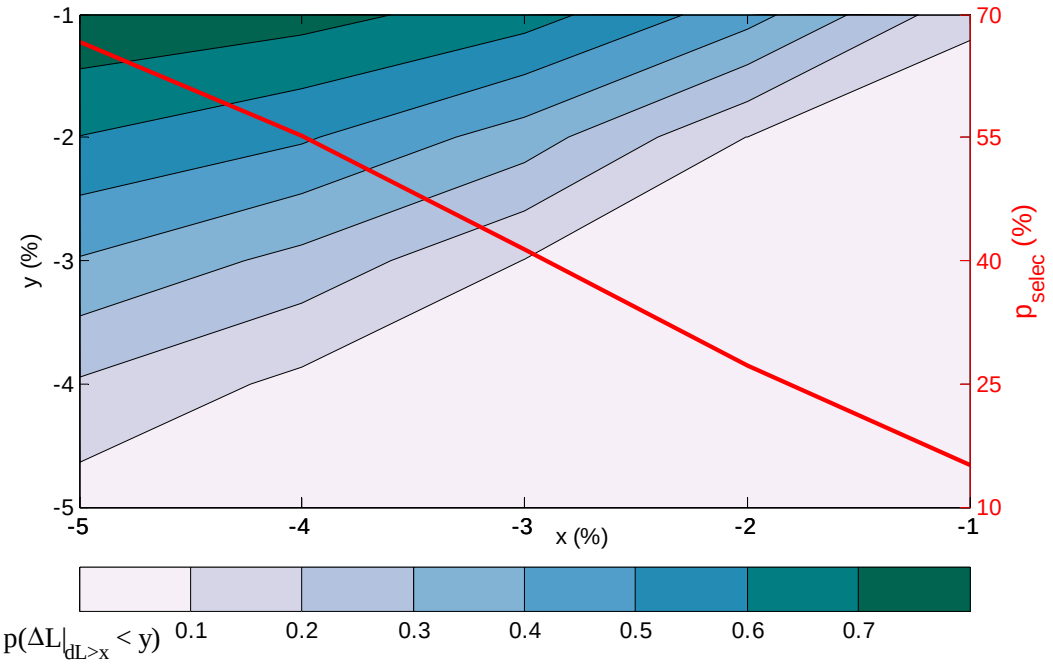


Figura 51 - Probabilidade $p(\Delta L_{dL>x} < y)$ de que sinais selecionados de acordo com o critério dL>x apresentem, como resultado da ação do RTO, valores de ΔL menores do que o valor y. Em vermelho: probabilidade (p_{selec}) de que uma janela de dados da classe $\mathscr{C}$ seja aceita como adequado para execução do RTO em função do valor de corte, x.

Em termos do uso, o modo mais natural do emprego do método seria o de operar com informações que o usuário forneça a partir do entendimento direto sobre a natureza probabilística dos resultados produzidos pelo RTO, fato que não está diretamente formulado na parametrização do método, embora esta natureza estocástica seja evidente a partir da observação das Figuras 48-51. Uma formulação mais conveniente da parametrização requerida do usuário seria a resposta à seguinte pergunta: Para dado nível de significância, qual o máximo desvio relativo do lucro que se deseja que seja produzido pelo RTO? Colocado sob esta forma a definição da parametrização do método fica diretamente vinculada a uma expectativa conveniente de desempenho,

diferentemente do que ocorria nos métodos convencionais. A formulação matemática equivalente à pergunta a ser respondida pelo usuário ao parametrizar o método é expressa na Equação (308). Note-se que foi necessário criar a função *xx*, que ajuda a vincular o desvio máximo admitido pelo usuário (valor de *x*) ao nível de significância associado aos valores de ΔL que ultrapassariam este limite caso o método seja utilizado (valor de α). Os sinais contidos na janela de dados serão aceitos para um determinado ciclo do RTO caso a condição (309) seja atendida.

$$xx = xx(x,\alpha) = \left\{ y \mid \int_{-\infty}^{y} p(\Delta L|_{dL>x})\, d\Delta L = \alpha \right\} \tag{308}$$

Aceitação: $dL > xx$ (309)

Para o presente caso, a relação *xx* foi obtida a partir do conjunto dos dados produzidos a partir do elenco de sinais participantes da classe $\mathscr{C}$ e não usados na formulação do método, ou seja, que não fizeram parte da solução de (298), que configurou o método. A dependência de *xx* com *x* e α mostrou-se aproximadamente linear, de modo que a relação $xx = a_0 + a_1.x + a_2.\alpha$ mostrou-se adequada para as condições estudadas, fazendo com que o percentil de ΔL correspondente a α fosse encontrado com erro de cerca de 5%. Deve-se notar, contudo, que este modo de formular uma parametrização conveniente para o problema possui um viés mais pragmático do que rigoroso, uma vez que a formulação do atendimento às condições do percentil da distribuição de ΔL deveria estar contida dentro do problema formulado em (298), o que o transformaria em um problema de programação estocástica não linear [31], condicionando sua solução a um nível de complexidade que tornaria extremamente custosa sua resolução computacional.

3.3. Adaptação do Modelo

Com o objetivo de apresentar a solução típica do problema de adaptação descrito na Equação (271), é necessário traçar o caminho das hipóteses que o fundamentam. Estas hipóteses restringem os problemas a serem resolvidos àqueles que possuem os requisitos elencados pelas Equações (310-314) [106,107].

R1) o processo é corretamente representado

$$\boldsymbol{fm} = \boldsymbol{f} \tag{310}$$

R2) os erros são aditivos às variáveis

$$\mathbf{Za} = \mathbf{Z} + \boldsymbol{\varepsilon} \tag{311}$$

R3) os erros apresentam média nula

$$\mu(\boldsymbol{\varepsilon}) = \mathbf{0} \tag{312}$$

R4) os erros seguem distribuição normal multivariável

$$\psi(\boldsymbol{\varepsilon}) \sim N(0, \mathbf{V}_{\varepsilon}) \tag{313}$$

R5) as variáveis necessárias (independentes) são isentas de corrupção:

$$\boldsymbol{\varepsilon}(\mathbf{in}) = \mathbf{0} \tag{314}$$

R6) os erros são independentes entre si

$$\text{covar}(e_i, e_j) = 0 \quad \forall\, i \neq j \tag{315}$$

Se os requisitos forem atendidos, a aplicação do método da máxima verossimilhança [51] resultará no problema de otimização apresentado na Equação (316), que representa a soma dos quadrados dos desvios entre as observações diretas e indiretas. Note-se que estes desvios são calculados ao longo da janela de dados, definidas pelas Equações (317-320), e escolhida de modo a prover informação suficiente para permitir a estimação de $\boldsymbol{\Theta}_j(\mathbf{upd})$ com o requerido nível de incerteza.

$$\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\min}\left(\Delta\mathbf{Z}^{\mathrm{T}}\,\Delta\mathbf{Z}\right)$$

$$s.a \qquad (316)$$

$$\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \end{cases}$$

Onde as janelas de dados obtidos por observação direta e indireta são, respectivamente:

$$\left[\mathbf{Za}(\mathbf{obj})\right]_{j,N,Tam} = \begin{bmatrix} \mathbf{Za}_{j-N+1}(\mathbf{obj}(1)) & \ldots & \mathbf{Za}_{j}(\mathbf{obj}(1)) \\ \ldots & \ldots & \ldots \\ \mathbf{Za}_{j-N+1}(\mathbf{obj}(\dim(\mathbf{obj}))) & \ldots & \mathbf{Za}_{j}(\mathbf{obj}(\dim(\mathbf{obj}))) \end{bmatrix} \qquad (317)$$

$$\left[\mathbf{Zm}(\mathbf{obj})\right]_{j,N,Tam} = \begin{bmatrix} \mathbf{Zm}_{j-N+1}(\mathbf{obj}(1)) & \ldots & \mathbf{Zm}_{j}(\mathbf{obj}(1)) \\ \ldots & \ldots & \ldots \\ \mathbf{Zm}_{j-N+1}(\mathbf{obj}(\dim(\mathbf{obj}))) & \ldots & \mathbf{Zm}_{j}(\mathbf{obj}(\dim(\mathbf{obj}))) \end{bmatrix} \qquad (318)$$

os desvios ao longo da janela de dados são expressos por:

$$\boldsymbol{a} = \left[\mathbf{Za}(\mathbf{obj})\right]_{j,N,Tam} - \left[\mathbf{Zm}(\mathbf{obj})\right]_{j,N,Tam} \qquad (319)$$

$$\Delta\mathbf{Z} = \begin{bmatrix} \boldsymbol{a}_{\bullet,1} \\ \boldsymbol{a}_{\bullet,2} \\ \ldots \\ \boldsymbol{a}_{\bullet,N} \end{bmatrix} = \begin{bmatrix} \mathbf{Za}_{n1}(i_1) - \mathbf{Zm}_{n1}(i_1) \\ \ldots \\ \mathbf{Za}_{n1}(i_I) - \mathbf{Zm}_{n1}(i_I) \\ \mathbf{Za}_{n2}(i_1) - \mathbf{Zm}_{n2}(i_1) \\ \ldots \\ \mathbf{Za}_{N}(i_I) - \mathbf{Zm}_{N}(i_I) \end{bmatrix} = \begin{bmatrix} \Delta\mathbf{Z}_1 \\ \ldots \\ \Delta\mathbf{Z}_I \\ \Delta\mathbf{Z}_{I+1} \\ \ldots \\ \Delta\mathbf{Z}_{N.I} \end{bmatrix} \qquad (320)$$

para simplificar a representação, foram usadas as variáveis:

nx = j-N+x , para indicar o instante no tempo

ix = **obj**(x), para indicar a variável na função objetivo, sendo I = dim(**obj**)

Note-se que o vetor $\Delta\mathbf{Z}$ tem dimensão N.I = N.dim(**obj**), com N igual ao número de amostragens de cada sinal na janela de dados, como anteriormente convencionado.

Pode ser mostrado [108] que, sendo atendidos R1 a R6, o estimador de mínimos quadrados apresenta características ótimas frente a outros estimadores. Traduzindo as propriedades dos estimadores calculados de acordo com (316), sob a vigência de R1-R6, pode-se garantir que o processo de adaptação incorpora as seguintes propriedades ao procedimento do RTO descrito na Figura 4:

- o estimador não acrescenta *bias* à adpatação

$$E\left(\mathbf{Zm}_j(\mathbf{upd})\right) = \mathbf{Z}_j(\mathbf{upd}) \tag{321}$$

- o estimador torna a adaptação *consistente* quando opera sobre $\left[\mathbf{Za}(\mathbf{obj})\right]_{j,N,Tam}$:

$$\lim_{N\to\infty} \; p\left(\left|\mathbf{Zm}_j(\mathbf{upd}(i)) - \mathbf{Z}_j(\mathbf{upd}(i))\right| < a \right) = 1, \;\; \forall a > 0, \; i = 1,2,...\dim(\mathbf{upd}) \tag{322}$$

- o estimador é eficiente, ou seja, possui a menor variância dentre quaisquer outras alternativas

Além destas características, caso as relações funcionais ***fm*** sejam lineares nos parâmetros estimados e caso inexistam restrições ativas, pode ser mostrado [51,106,107] que a região que contém a expectativa, sob dado nível de significância, dos parâmetros estimados com a Equação (316), é delimitada por um hiper-elipsóide cujo centro coincide com os valores reais do processo. Portanto, sob estas condições, as estimativas $\mathbf{\Theta}_j(\mathbf{upd})$ serão variáveis estocásticas cujos valores estarão contidos, no espaço dos parâmetros estimados, no interior de uma região cuja superfície pode ser convenientemente descrita de forma analitica. A possibilidade de formular a expressão genérica do contorno da região de confiança e de prever os efeitos das condições do problema na qualidade das estimativas é muito conveniente e tem sido explorada [24, 25], uma vez que permite adicionar previsibilidade aos efeitos sobre a camada de otimização econômica do RTO. Contudo, para o problema não linear não há previsibilidade sobre a forma e a dimensão da região de confiança, que pode mesmo ser aberta [106], inviabilizando a previsão unificada da expectativa das consequências do processo de estimação sobre o resultado global do RTO.

Caso os erros não sejam independentes entre si, ou seja, se o requisito R6 (Equação 315) não for atendido, a função objetivo em (316) pode ser modificada [107]

de modo a transformar-se na Equação (323), marcada pela inclusão do termo correspondente à inversa da matriz variância-covariância **V** dos erros (324). O uso da Equação (323) também se aplica caso os erros sejam independentes mas não possuam variância constante ao longo das observações (heteroscedasticidade), sendo que os termos fora da diagonal de **V** serão nulos para esta situação.

$$\begin{aligned} &\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\min}\left(\Delta\mathbf{Z}^{\mathrm{T}}\,\mathbf{V}^{-1}\,\Delta\mathbf{Z}\right) \\ &s.a \\ &\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \end{cases} \end{aligned} \tag{323}$$

A matriz de variância-covariâncias é dada por:

$$\mathbf{V} = \begin{bmatrix} \mathrm{var}(\Delta\mathbf{Z}_1) & \mathrm{covar}(\Delta\mathbf{Z}_1,\Delta\mathbf{Z}_2) & \ldots & \mathrm{covar}(\Delta\mathbf{Z}_1,\Delta\mathbf{Z}_{N.I}) \\ \mathrm{covar}(\Delta\mathbf{Z}_2,\Delta\mathbf{Z}_1) & \mathrm{var}(\Delta\mathbf{Z}_2) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{covar}(\Delta\mathbf{Z}_{N.I},\Delta\mathbf{Z}_1) & \ldots & \ldots & \mathrm{var}(\Delta\mathbf{Z}_{N.I}) \end{bmatrix} \tag{324}$$

Note-se que, para a janela de dados que contém N pontos de amostragem por variável, a dimensão da matriz **V** é (**N**.dim(**obj**))^2, o que traz considerável dificuldade de ordem prática para a correta avaliação de seus elementos. No exemplo a seguir será apresentado um caso simples que mostra as consequências da negligência em relação à avaliação de **V**, assim como dos efeitos da não linearidade sobre as expectativas dos parâmetros estimados.

O não atendimento dos requisitos pode retirar as propriedades ótimas do estimador de mínimos quadrados. Com toda certeza o requisito mais difícil de ser atendido, ainda que seja o de ocorrência mais provável, é o da representação perfeita do modelo do processo (R1). Outro requisito cujo descumprimento é muito comum mas cuja importância e consequencia costumam ser subestimadas e negligenciadas é o que prevê que as variáveis independentes devam ser medidas de forma perfeita (R5), sendo difícil de ser contornado sob a estrutura do estimador de mínimos quadrados. Mesmo que sob a vigência de todos os demais requisitos, ele coloca alguns sérios obstáculos à manutenção das propriedades desejáveis do estimador [109].

Para ilustrar este caso específico será apresentado um exemplo muito simples, do processo linear nos parâmetros apresentado na Equação (325). Se todos os requisitos forem atendidos e se o único sinal corrompido for ya (versão de y adquirida pela instrumentação), a estimativa do parâmetro α se dará pela solução do problema de otimização (326,327), que pode ser analiticamente resolvido para este caso simples por meio do atendimento da condição (328). Deste modo, o valor de α obtido pelo estimador de mínimos quadrados é dado pela Equação (329).

Como o termo do denominador na Equação (329) é determinístico, é fácil perceber a validade das relações enunciadas na Equação (330): sob a prevalência de erro gaussiano de média nula em y e se o modelo for linear nos parâmetros estimados, o estimador de mínimos quadrados produzirá parâmetros cuja distribuição será gaussiana cuja média coincide com o valor real.

$$y = \alpha x \tag{325}$$

$$\hat{\alpha} = \arg\min_{\hat{\alpha}} \left(\sum_{i=1}^{N} (ya_i - ym_i)^2 \right) = \arg\min_{\hat{\alpha}} \left(\sum_{i=1}^{N} (y_i + \varepsilon_i - ym_i)^2 \right) \tag{326}$$

$$\hat{\alpha} = \arg\min_{\hat{\alpha}} \left(\underbrace{(\alpha x_1 + \varepsilon_1 - \hat{\alpha} x_1)^2 + (\alpha x_2 + \varepsilon_2 - \hat{\alpha} x_2)^2 + \ldots + (\alpha x_N + \varepsilon_N - \hat{\alpha} x_N)^2}_{S} \right) \tag{327}$$

$$\hat{\alpha} : \frac{\partial S}{\partial \hat{\alpha}} = 0 \tag{328}$$

$$\hat{\alpha} = \alpha + \frac{\sum_{i=1}^{N} x_i \varepsilon_i}{\sum_{i=1}^{N} x_i^2} \tag{329}$$

$$\varepsilon \sim N(0, \sigma) \;\Rightarrow\; E(\hat{\alpha}) = \alpha;\; \hat{\alpha} \sim N(\alpha, \sigma_\alpha) \tag{330}$$

Contudo, as propriedades expressas na Equação (330) não serão esperadas se a variável necessária (independente) x estiver contaminada pelo ruído gaussiano ε, *ainda*

que y seja medida de forma fidedigna. Neste caso, o valor estimado de α é obtido via o problema de otimização expresso em (331), cuja solução analítica é mostrada na Equação (332). Note-se que sob a presença de erros na variável independente não há garantia de que o estimador de mínimos quadrados seja consistente ou eficiente [109].

$$\hat{\alpha} = \arg\min_{\hat{\alpha}} \left(\sum_{i=1}^{N} (ya_i - ym_i)^2 \right) = \arg\min_{\hat{\alpha}} \left(\sum_{i=1}^{N} \left(y_i - \hat{\alpha}(x_i + \varepsilon_i) \right)^2 \right) \tag{331}$$

$$\hat{\alpha} = \alpha \left(\frac{\sum_{i=1}^{N} (x_i^2 + \varepsilon_i x_i)}{\sum_{i=1}^{N} (x_i^2 + 2\varepsilon_i x_i + \varepsilon_i^2)} \right) \tag{332}$$

Na Figura (52) estão apresentadas as estimativas das funções densidade de probabilidade do *bias* da estimativa de α. Neste exemplo, os valores de x consistem de 5 pontos igualmente espaçados no intervalo [3 5], tendo sido usadas 1.10^6 estimativas para simular as pdf's. Como observado, sob as condições apresentadas, as estimativas possuem *bias* e α não apresenta distribuição gaussiana mesmo se o problema for linear nos parâmetros e se ε apresentar distribuição gaussiana. Também se pode observar a dependência do *bias* com o desvio-padrão do ruído adicionado. É um fato conhecido [110] que a presença de erro na variável dependente faz com que a inclinação da reta seja sempre subestimada e tenda a zero à medida que a variabilidade do erro aumenta. A expressão do valor esperado do bias quando o modelo é expresso pela equação de uma reta pode ser analiticamente descrito [111] como nas Equações (333,334), onde σ_x e σ_ε são, respectivamente, o desvio padrão da variável independente e da variável dependente e N é o número de pontos amostrados.

$$E(\alpha - \hat{\alpha}) = \alpha \frac{r(\rho + r)}{1 + 2\rho r + r^2} \tag{333}$$

$$\begin{cases} \rho = \sigma_{x\varepsilon} / (\sigma_x \sigma_\varepsilon) \\ r = \sigma_\varepsilon / \sigma_x \\ \sigma_x^2 = \sum_{i=1}^{N} (x_i - \bar{x}_i)^2 / N \\ \sigma_{x\varepsilon} = \text{covar}(x, \varepsilon) \end{cases} \tag{334}$$

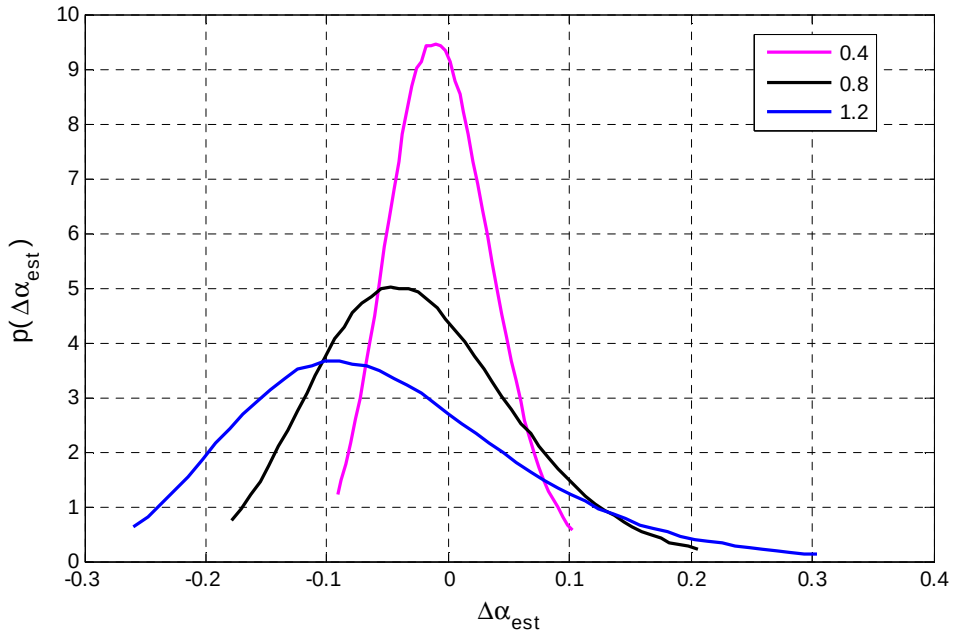


Figura 52 – Função distribuição de probabilidade do bias de estimativa para problema linear com erro na variável independente, x. Legenda: desvio-padrão de **ε**.

Um segundo exemplo será mostrado em seguida para realçar os efeitos na não linearidade, da ocorrência de erros na medição das variáveis necessárias e da falha em caracterizar V quando a restrição de independência entre as variáveis for violada. Este exemplo será apresentado utilizando de forma didática a representação proposta no Capítulo 2:

- <u>Descrição do processo e das condições de operação</u>:

Consideremos o processo descrito pela seguinte relação funcional:

$$\boldsymbol{f}: \; y = c_1 \alpha + c_2 \frac{\beta x}{c_3 + \alpha x} + c_3 \beta x \tag{335}$$

cujo elenco de entidades matemáticas está agrupado em Z:

$$\{\mathbf{Z}\} = \{x, \alpha, \beta, c_1, c_2, c_3, y\} \tag{336}$$

sendo que apenas variáveis x e y estão sujeitas a mudanças:

$$\mathbf{var} = [1,\ 7] \ \Rightarrow\ \mathbf{fix} = [2,\ 3,\ 4,\ 5,\ 6]; \tag{337}$$

com os valores imutáveis definidos como:

$$\mathbf{Z(fix)} = [2.9\ 1.7\ \ 0.3\ \ 2.1\ \ 0.4]^T \tag{338}$$

- <u>Representação do processo e configuração do RTO:</u>

A representação disponível do processo mantém a integridade das informações:

$$\begin{cases} \boldsymbol{fm} = \boldsymbol{f} \\ \boldsymbol{Zm_0} = \mathbf{Z}_0 \end{cases} \tag{339}$$

As variáveis x e y são obtidas por observação direta:

$$\mathbf{ms} = [1\ \ 7] \tag{340}$$

A etapa de adaptação do RTO aplicado a este processo está configurada da seguinte forma:

o elenco das variáveis necessárias $x, \alpha, \beta, c_1, c_2, c_3$:

in = [1 2 3 4 5 6] ⇒ **out** = [7]

c_1, c_2 e c_3 escolhidos para ficarem inalterados ao longo dos ciclos do RTO

atr = [4 5 6]

α e β são graus de liberdade do processo de adaptação

upd = [2 3]

a variável y faz parte da função objetivo de adaptação

obj = [7]

a medição da variável x é incorporada ao modelo do processo (70):

$$\mathbf{ms}^{-} = [1]$$

- Relações funcionais aplicáveis à janela de dados:

Supondo que uma janela de dados com N = 25 pontos seja utilizada:

$$[\mathbf{Z}]_{1,25,Tam} = [\mathbf{Z}_1 \ ... \ \mathbf{Z}_{25}]$$

O estímulo consiste na variação linear de x:

$$\begin{aligned} &\mathbf{Z}_1(1) = x_{ini} \\ &\mathbf{Z}_i(1) = \mathbf{Z}_{i-1}(1) + (x_{fim} - x_{ini})/25, \quad i > 1 \end{aligned} \tag{341}$$

De (335) e (336) vêm as relações funcionais do processo:

$$[\,\boldsymbol{f}\,]_{1,25} : \begin{cases} \mathbf{Z}_1(7) = \mathbf{Z}_1(4)\mathbf{Z}_1(2) + \mathbf{Z}_1(5)\dfrac{\mathbf{Z}_1(3)\mathbf{Z}_1(1)}{\mathbf{Z}_1(6) + \mathbf{Z}_1(2)\mathbf{Z}_1(1)} + \mathbf{Z}_1(6)\mathbf{Z}_1(3)\mathbf{Z}_1(1) \\ ... \\ \mathbf{Z}_{25}(7) = \mathbf{Z}_{25}(4)\mathbf{Z}_{25}(2) + \mathbf{Z}_{25}(5)\dfrac{\mathbf{Z}_{25}(3)\mathbf{Z}_{25}(1)}{\mathbf{Z}_{25}(6) + \mathbf{Z}_{25}(2)\mathbf{Z}_{25}(1)} + \mathbf{Z}_{25}(6)\mathbf{Z}_{25}(3)\mathbf{Z}_{25}(1) \end{cases} \tag{342}$$

Funções de atribuição (vide Equação (54)):

$$[\,\boldsymbol{f}_{atr}\,]_{1,25} : \begin{cases} \begin{cases} \mathbf{Zm}_1(4) = 0{,}3 \\ \mathbf{Zm}_1(5) = 2{,}1 \\ \mathbf{Zm}_1(6) = 0{,}4 \end{cases} \\ ... \\ \begin{cases} \mathbf{Zm}_{25}(4) = 0{,}3 \\ \mathbf{Zm}_{25}(5) = 2{,}1 \\ \mathbf{Zm}_{25}(6) = 0{,}4 \end{cases} \end{cases} \tag{343}$$

Funções de medição (vide (53)):

$$[\,\boldsymbol{f}_{med}\,]_{1,25}:\begin{cases}\{\mathbf{Zm}_1(1)=\mathbf{Za}_1(1)\\ \ldots\\ \{\mathbf{Zm}_{25}(1)=\mathbf{Za}_{25}(1)\end{cases} \tag{344}$$

Devem ser adicionadas relações de continuidade dos valores das variáveis **upd**, uma vez que os valores das variáveis estimadas são constantes ao longo de toda a janela de dados:

$$[\,\boldsymbol{f}_{cont}\,]_{1,25}:\begin{cases}\begin{cases}\mathbf{Zm}_2(2)=\mathbf{Zm}_1(2)\\ \mathbf{Zm}_2(3)=\mathbf{Zm}_1(3)\end{cases}\\ \ldots\\ \begin{cases}\mathbf{Zm}_{25}(2)=\mathbf{Zm}_1(2)\\ \mathbf{Zm}_{25}(3)=\mathbf{Zm}_1(3)\end{cases}\end{cases} \tag{345}$$

O conjunto de Equações (342-345) define os dois graus de liberdade disponíveis para o problema de adaptação, conforme mostrado na Equação (346).

$$GL=\dim([\mathbf{Z}]_{1,25})-\dim([\boldsymbol{f}\,]_{1,25})-\dim([\boldsymbol{f}_{atr}\,]_{1,25})\;-\dim([\boldsymbol{f}_{med}\,]_{1,25})-\dim([\boldsymbol{f}_{cont}\,]_{1,25})$$

$$GL=N\dim(Z)-N\dim(\boldsymbol{fm})-N\dim(\mathbf{atr})\;-N\dim(\boldsymbol{ms}^{-})-(N-1)\dim(\mathbf{upd}) \tag{346}$$

GL = 2

Uma vez caracterizado o problema, serão apresentados dois casos que diferem pela conformação dos erros associados às medições a partir da estrutura geral apresentada na Equação (347). Como apresentado na Tabela 9, para o caso A os requisitos R1-R6 são atendidos, enquanto que para o caso B a variável necessária, x, não está isenta de corrupção, assim como não há independência dos erros, violando, respectivamente, as restrições R5 e R6.

Deste modo, definindo-se x_{ini}=0, x_{fim}=5 como a faixa de variação do estímulo x, conforme a Equação (341), e fazendo σ = 1/15, pode-se observar as expectativas dos erros relativos dos parâmetros estimados na Figura 53 geradas a partir de $1x10^5$ simulações.

$$\varepsilon(n)=\phi\varepsilon(n-1)+\delta(n);\qquad \delta\sim N(0,\sigma) \tag{347}$$

Tabela 9 – Parâmetros de caracterização da estrutura dos erros para os casos A e B

Casos	ϕ	σ_x	σ_y
A	0	0	0.05
B	0.3	0.048	0.048

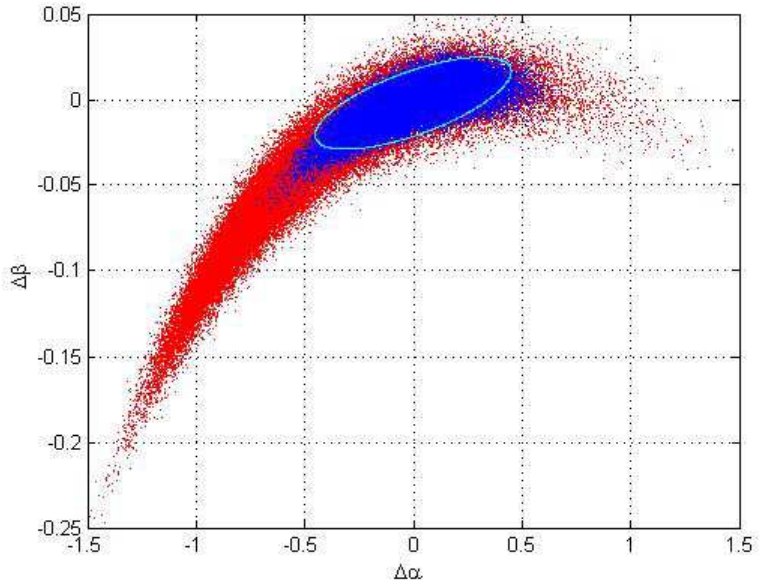


Figura 53 – Desvio esperado entre as estimativas de α e β e seus valores verdadeiros. Pontos azuis: caso A; Pontos Vermelhos: caso B. elipse: região de 95% de confiança para o modelo linearizado.

O fato de o modelo ***f***, na Equação (335), ser não linear nos parêmetros estimados α e β, não permite que se suponha que as regiões de confiança das estimativas no espaço dos parâmetros estimados sejam representadas por hiper-elipsóides. Ainda que todos os requisitos R1-R6 sejam atendidos, como no caso A, não será possível descrever um formato universal para a região de confiança, o que pode ser comprovado por meio da Figura 53.

No caso B dois requisitos não são atendidos: há a presença de erros nas variáveis necessárias do modelo; existe um padrão de auto-correlação não identificado nos erros, sendo que o não atendimento destes dois requisitos é muito comum em casos reais. Contudo, apesar destas ocorrências serem comuns, suas consequências não são menos dramáticas, como será visto a seguir.

Com esta configuração do problema, as principais propriedades do estimador de mínimos quadrados são perdidas. As estimativas não possuem mais a garantia teórica de *bias* nulo (Equação 321). No presente caso, o desvio relativo médio de α é de -13%, sendo de -2% para β. Além disso, o volume ocupado pela região de confiança aumenta significativamente, conforme observado na Figura 53. Outra consequência é que, se alguns dos requisitos R1-R6 não forem atendidos, o estimador deixa de ser consistente, conforme definido pela propriedade (322), esperada quando o tamanho da janela de dados cresce de forma ilimitada. Como visto na Figura 54, o comportamento assintótico das estimativas se dirige a um valor diferente do verdadeiro. Se a Figura 54 for comparada com a Figura 55, que traduz o caso A, pode-se notar a grande diferença que o atendimento aos requisitos pode fazer em termos de acurácia e consistência do estimador.

Além destes fatos, é importante ressaltar que, para problemas não lineares nos parâmetros, o formato da região que descreve as expectativas das estimativas pode ser profundamente alterado em função do domínio dos valores das variáveis de entrada do modelo. A alteração dos valores que definem os extremos da faixa de variação de x (x_{ini} e x_{fim} em (341)) traz as consequências sobre as regiões esperadas para os valores estimados conforme mostrado na Figura 56.

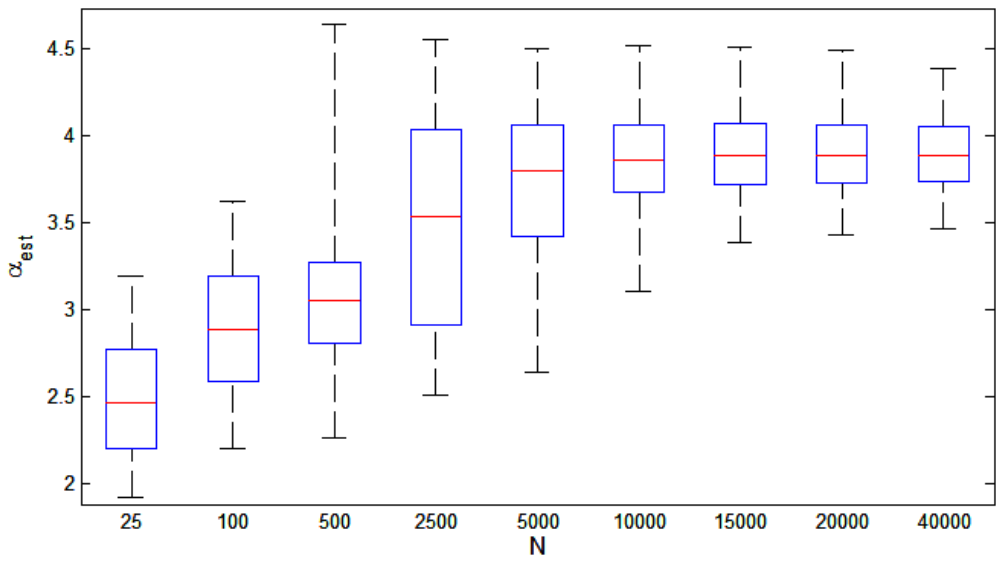


Figura 54 – Distribuição das estimativas de α em função do tamanho da janela de dados para o caso B. Valor real de α = 2,9.

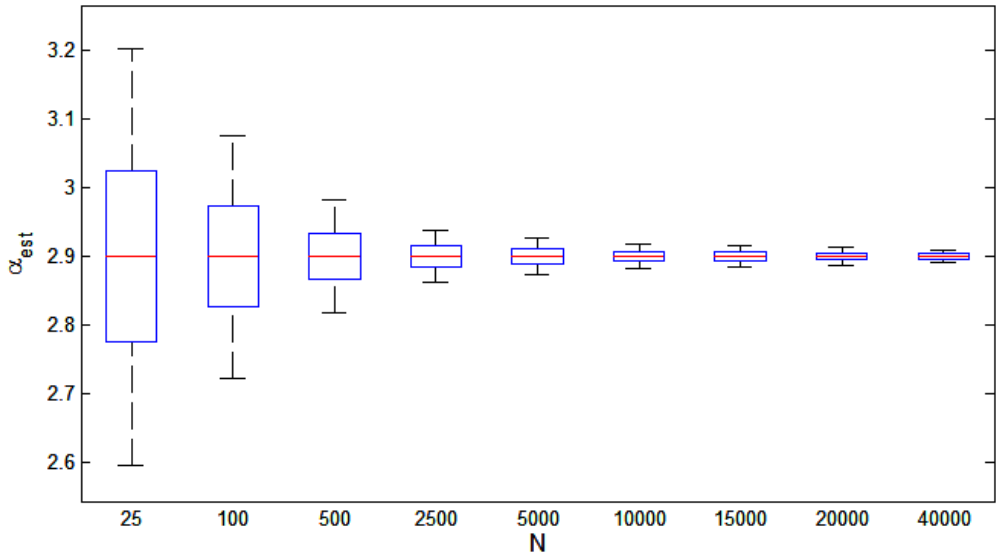


Figura 55 - Distribuição das estimativas de α em função do tamanho da janela de dados para o caso A. Valor real de α = 2,9.

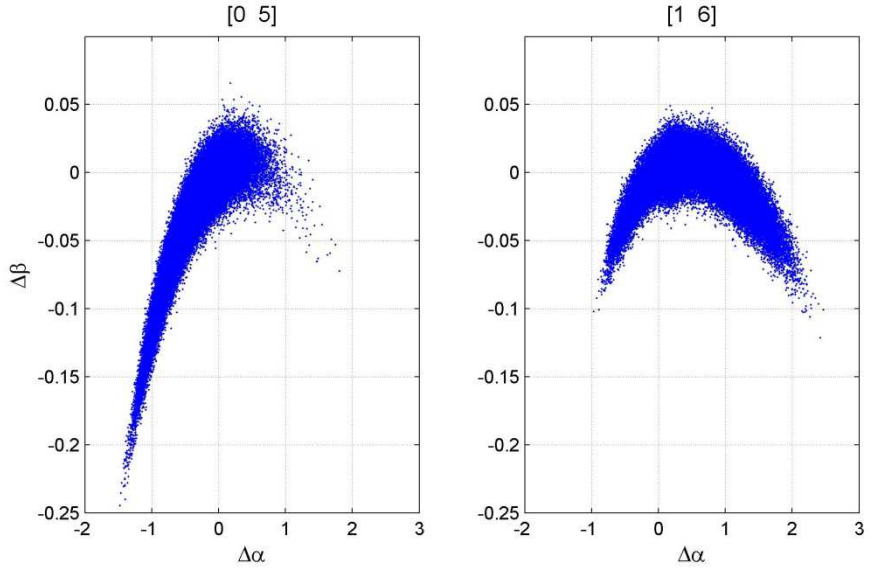


Figura 56 – Desvio das estimativas de α e β em função do domínio dos valores de x para o caso B. Intervalo de valores de x indicado no topo de cada gráfico.

3.3.1. O Papel da Janela de Dados

A adaptação do modelo é alimentada pela informação acumulada ao longo do horizonte pregresso contido na janela de dados (Equação 317). O uso desta estratégia está relacionado a um contexto de adaptação que faça uso de um estimador consistente, cujas propriedades assintóticas permitam supor a vinculação entre a qualidade das estimativas e o tamanho da janela de dados. Isto pode ser facilmente percebido da relação entre N, o número de ponto na janela de dados, e o desvio padrão da estimativa de α, conforme a Equação (329) do exemplo da Seção 3.3. Como mostrado na Figura 57, a consistência do estimador garante que a variabilidade do parâmetro estimado seja tão pequena quanto se queira, bastando para isto manipular-se o tamanho da janela de dados.

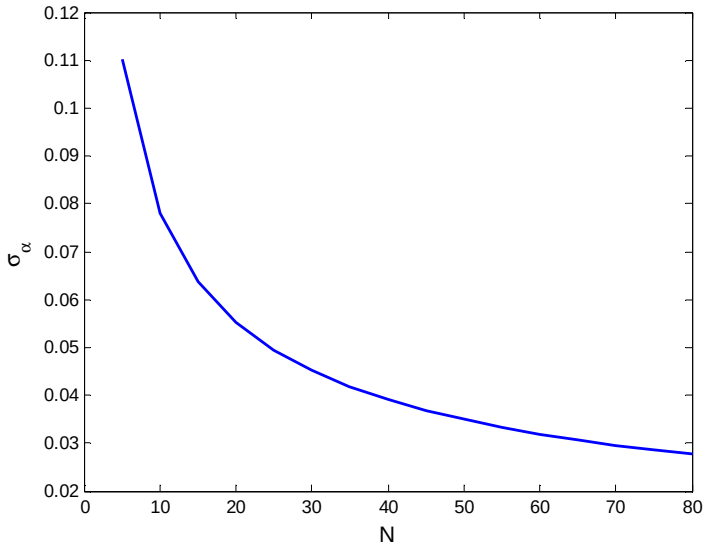


Figura 57 – Efeito do tamanho da janela de dados no desvio-padrão da estimativa de α.

Se o efeito visto na Figura 57, fruto do uso de um estimador consistente, fosse válido em um caso de aplicação real, seria muito fácil obter estimativas “perfeitas”, bastando para isto que a instrumentação adquirisse os sinais com uma elevada frequência de amostragem. O fato de esta estratégia não garantir tal sucesso em um sistema real de RTO diz muito a respeito da natureza dos sinais envolvidos. A hipótese de que toda a corrupção dos sinais provém de erros aditivos aleatórios não correlacionados não se

sustenta face à incapacidade desta estratégia garantir melhoria ilimitada nas estimativas. Na verdade, o insucesso desta estratégia na prática é que serve para provar que os sinais reais são sujeitos à correlação do sinal de ruído além de mostrar a presença de informações não medidas, Z(**dum**).

Como anteriormente ressaltado, o uso de uma janela de dados no processo de adaptação está intrinsicamente vinculado à expectativa de consistência do estimador, o que induziria a tendência de uma janela a mais extensa possível. Por outro lado, a expectativa de variabilidade dos parâmetros cujos valores estão sendo estimados representa um contraponto a esta estratégia, na medida em que o parâmetro estimado é suposto invariável ao longo da janela de dados (vide exemplo na Equação (345)). Para que estes fatores sejam corretamente ponderados é necessário que se possua conhecimento sobre:

- a natureza da corrupção, $\psi(\boldsymbol{\varepsilon})$, para discernir a relação de compromisso entre aumento da janela e diminuição da variabilidade das estimativas

- a expectativa da variabilidade do parâmetro a ser estimado, para evitar os efeitos de *aliasing* devido à subamostragem inerente ao uso de uma janela de dados.

Para esclarecer melhor este último ponto, note-se que, para o caso em que os sinais sejam perfeitamente adquiridos, $p(\mathbf{ZZa}(\mathbf{ms}) = \mathbf{ZZ}(\mathbf{ms}))=1$, em tese não há necessidade de se usar uma janela de dados. A estimativa pode ser feita a cada ponto adquirido desde que o sistema possua graus de liberdade para tal. Contudo, a discretização associada ao processo de amostragem realizado pela instrumentação, como visto em (48), impõe que a frequência de amostragem do sistema de aquisição seja no mínimo o dobro da máxima frequência componente de cada sinal (Equação 348), de acordo com o teorema de Shannon-Nyquist [112]. A máxima frequência componente de cada variável em **ZZ** ao longo do tempo é aquela a partir da qual o espectro de potência é nulo, como expresso nas Equações (349,350).

$$\frac{1}{\underbrace{t_{j+1}-t_j}_{Tam}} > 2\,\omega_{\max,k}\left(\mathbf{ZZ}(k)\right), \quad k=1..\dim(\mathbf{upd}) \tag{348}$$

$$\omega_{\max,k} = \left\{ \omega \middle| \left|X(ix)\right|^2 = 0, \ \forall x > \omega \right\} \tag{349}$$

$$X(\omega i) = \left\langle ZZ_t(k),\ e^{-i\omega t} \right\rangle \tag{350}$$

Já para o caso mais geral em que os sinais são corrompidos, $p\left(\mathbf{ZZa}(\mathbf{ms}) = \mathbf{ZZ}(\mathbf{ms})\right)<1$, é usada uma janela de dados com a finalidade de diminuir a incerteza associada às estimativas. Como é suposta a constância do valor estimado ao longo da janela de dados (vide Equação 345), a frequência de amostragem das variáveis estimadas está agora inversamente relacionada ao intervalo de tempo abarcado pela janela de dados, e não mais ao intervalo com que o sinal é adquirido pelo instrumentação. Deste modo, a precaução a ser tomada diz respeito à adequação entre a máxima frequência de variação do parâmetro real e a frequência associada à largura temporal da janela de dados:

$$\frac{1}{\underbrace{t_j - t_{j-n+1}}_{nT_{am}}} > 2\omega_{\max,k}\left(\mathbf{ZZ}(\mathbf{upd}(k))\right), \quad k=1..\dim(\mathbf{upd}),\ \textit{dado}\ \left[\mathbf{Za}\right]_{j,n,Tam} \tag{351}$$

Note-se que quando foram enumeradas as possíveis configurações de um sistema de RTO, como mostrado em (125), não havia diferenciações prévias entre as combinações elencadas. Contudo, a Equação (351) traz consigo uma importante diferenciação entre as variáveis **upd** e as demais variáveis **in** e **out**. Enquanto estas admitem uma frequencia máxima limitada a $0{,}5/T_{am}$, as variáveis adaptadas devem possuir uma largura de banda menor que $0{,}5/(nT_{am})$, onde n é o número de amostras contidas na janela de dados. É interessante notar que, se a frequência máxima de ao menos uma das variáveis adaptadas superar o limite suposto na Equação (351) isto fará com que o processo de estimativa produza valores corrompidos por *aliasing* [112] e que esta corrupção poderá ser disseminada para todas as demais variáveis **upd** em virtude da natureza multivariável do processo de adaptação.

Aproveitando o mesmo modelo y = αx apresentado na Equação (325), pode-se apresentar um exemplo no qual seja explorada a dicotomia entre a diminuição da incerteza e a corrupção da forma de onda associadas ao tamanho da janela de dados. A configuração do sistema de RTO associado a este modelo pode ser vista na Tabela 10.

Tabela 10 – Configuração do RTO usado como exemplo sobre a influência da janela de dados na adaptação do modelo

{Z}	in	ms	upd	fix	f (= fm)	T_{am}
{x α t ω y}	[1 2 3 4]	[1 3 5]	[2]	[4]	$\begin{cases} y=\alpha x \\ \alpha=1+sen(2\pi\omega t)/4 \end{cases}$	0,1 s

Note-se que o parâmetro estimado, α, varia ao longo do tempo com uma frequência dada por ω. A Figura 58 mostra o valor esperado do valor absoluto do *bias* de estimação (Equação 352). Pode-se perceber que, quando o nível de ruído é pequeno, predomina o efeito negativo da janela de dados sobre a corrupção da dinâmica do comportamento do parâmetro estimado, fazendo com que janelas maiores sempre causem maiores problemas qualquer que seja a frequência de variação de α. Para sinais mais ruidosos, o tamanho da janela influencia de modo mais variado. Janelas mais largas podem ser mais úteis do que janelas estreitas em determinada faixa de frequências de variação do parâmetro estimado em função da preponderância do efeito benéfico da diminuição da incerteza sobre a corrupção da dinâmica.

$$E\left(\left|\alpha-\alpha_{est}\right|\right)=E\left(\left|\alpha_j-\hat{\alpha}|_{[\mathbf{Z}]_{j,n,Tam}}\right|\right) \tag{352}$$

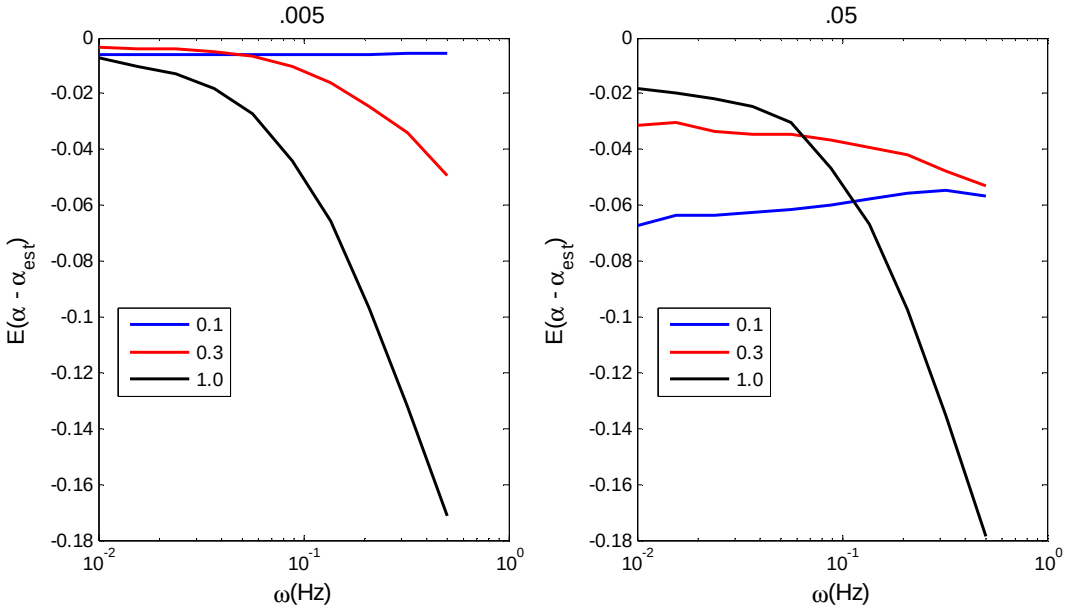


Figura 58 – Valor esperado do bias de estimação em função da frequência de variação do parâmetro estimado para o exemplo descrito no texto. Legenda: intervalo de tempo (s) abarcado pela janela de dados. No título de cada gráfico está indicado o desvio-padrão do ruído.

4. RTO Industrial

Nesta Seção serão descritos os modos de operação comuns a dois sistemas de Otimização em Tempo Real atualmente em uso em unidades industriais, o AspenPlus versão 7.1 , da Aspentech, e o Romeo versão 5.3, da Invensys, apresentados neste texto como os protótipos de implementações em plantas reais do RTO em duas camadas. Um exemplo típico de aplicação é mostrado na Figura 59, onde é apresentada uma implementação do Romeo em uma unidade de destilação atmosférica. Nela podem ser vistas três instâncias da implementação do RTO:

A) Detecção do estado estacionário (*Steady State Detection*, SSD), em que são avaliadas as informações das variáveis de processo adquiridas ao longo de uma janela de dados de modo a determinar se a planta está estável

B) Chaveamento da execução do RTO baseado em informações da estabilidade da unidade, da carga da unidade e do status do sistema RTO. A sinalização para o início de um novo ciclo de otimização também costuma respeitar um intervalo mínimo desde a execução anterior, tempo que tipicamente está na faixa de 30 minutos a duas horas para aplicações em destilação.

C) Execução sequencial da adaptação do modelo e da otimização da função de gerenciamento econômico da unidade.

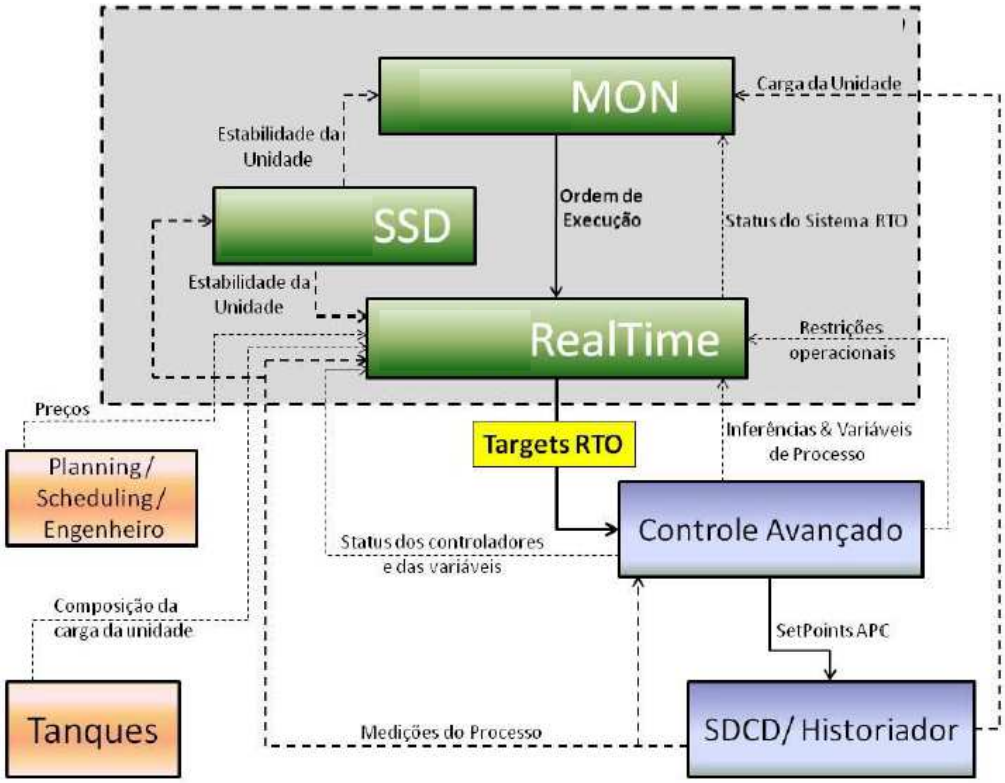


Figura 59 – Fluxograma do sistema de RTO da Invensys em uso em uma unidade de destilação de petróleo.

O RTO é integrado ao sistema de *scheduling*/planejamento da produção, do qual recebe as diretrizes a serem cumpridas; à logística de armazenamento, por onde é informada a composição dos tanques de carga; e ao SDCD e à base de dados do historiador de dados, dos quais recebe os valores das medições. As variáveis de decisão do RTO são implementadas no sistema de controle avançado, que cuidará da trajetória dinâmica percorrida pela unidade até a implementação completa das soluções do RTO.

No presente trabalho os dados de plantas reais referem-se a implementações de sistemas de RTO da Aspentench (AspenPlus) e da Invensys (Romeo) atualmente em uso em unidades de destilação de petróleo em distintas refinarias. O processo onde está implementado o AspenPlus será aquele preferencialmente usado para fins de análise dos dados produzidos, enquanto que o Romeo será oportunamente refenciado para fins de comparação da arquitetura e algoritmos.

4.1. Detecção de estacionariedade (SSD)

Dentro do conceito tradicional de implementação de RTO estático com otimização em duas etapas, ambas as empresas, Aspentech e Invensys, incorporam em seus sistemas um módulo de SSD cujo resultado subsidia a autorização de mais um ciclo de otimização do RTO. Os métodos utilizados são similares, seja em forma ou ao menos em essência, às abordagens apresentadas na Seção 3.2.2.

A Aspentech, no *software* Aspen Plus RTO, oferece duas opções de detecção, apresentadas ao usuário sob a denominação de "*Método Estatístico*" e de "*Método Heurístico*". O assim chamado método estatístico nada mais é do que o método da estatística R (Equação 224). Há, contudo, uma importante diferença de emprego, pois é disponibilizada ao usuário a opção de usar um parâmetro de sintonia. Na prática, este parâmetro, T, opera como mostrado na Equação (353).

$$\begin{aligned} &R = X / s^2; \quad X = \max(s_d^2, T) \\ &R > R_{critico} \iff \text{processo estacionar io} \end{aligned} \tag{353}$$

O assim chamado *método heurístico* faz uso de duas versões do sinal original, sujeitas a filtros passa-baixas com distintas frequências de corte, que são indiretamente manipuladas pelos parâmetros f_L e f_P, conforme a Equação (354). A versão com frequência de corte mais baixa é chamada de "filtro pesado", X_p, enquanto que a outra é chamada de "filtro leve", X_L. Se a diferença entre as duas versões for menor que determinada tolerância o sinal é considerado estacionário. No uso cotidiano este autor observou que este método é preterido em relação ao anterior. Provavelmente devido ao maior ônus imposto ao usuário, que deve definir um total de ($3.n_{variaveis}$) parâmetros, pois cada sinal requer um valor de f_L, f_P e de tolerância, com o agravante que as tolerâncias devem estar na unidade de cada sinal pois não são expressas em termos relativos.

$$\begin{aligned} \mathrm{X_L(i)} &= f_L \mathrm{X}(i) + (1 - f_L) X_L(i-1) \\ \mathrm{X_P(i)} &= f_P \mathrm{X}(i) + (1 - f_P) X_P(i-1) \end{aligned} \tag{354}$$

Deve-se notar o papel exercido pelos parâmetros de configuração deixados à escolha do usuário. No método "estatístico", o uso da tolerância T (Equação 353) permite que, na prática, o usuário defina por conta própria o que ele quer que seja

considerado como estacionário, e qualquer pretensão estatística advogada para si pelo método é anulada. No caso do método "heurístico", a natureza arbitrária dos parâmetros coloca o usuário no comando da decisão, direcionando as decisões segundo suas impressões arbitrárias, embora, dado o elevado número de escolhas, seja provável que o usuário não consiga antecipar os efeitos destas escolhas no formato dos sinais.

No RTO Romeo também são apresentadas duas opções de testes de estacionariedade. Uma delas é o mesmo teste baseado na estatística R (Equação 224), que também é oferecido pelo AspenPlus. A outra opção consiste no teste de hipósteses para decidir se os valores médios de duas metades da janela de dados do sinal são iguais. Inicialmente, é testada a hipótese de que as variâncias sejam iguais, o que é feito com o tradicional teste da razão entre as variâncias das duas metades do sinal por meio da estatística *F*, de Fisher [77]. Se as variâncias forem consideradas diferentes, o procedimento aplicado é idêntico ao método de Ackeman-Schladt, apresentado nas Equações (216,218). Em caso contrário, é realizado o teste para verificar se a diferença entre as médias é menor que dada tolerância *Tol*, o que é feito baseado no fato que a estatística *t* (355,356) supostamente apresenta distribuição de Student com n_1+n_2-2 graus de liberdade, sendo $n_1 = n_2$ o número de pontos em cada janela de comparação neste caso. No Romeo, estes testes consideram sempre um nível de significância $\alpha = 10\%$.

$$s = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n1+n2-2} \tag{355}$$

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right| - Tol}{\sqrt{s(1/n_1 + 1/n_2)}} \tag{356}$$

Em ambos os *softwares* cabe ao usuário decidir, além dos parâmetros do método de detecção de estacionariedade, o conjunto de sinais que será submetido ao teste de estacionariedade, **ee** em (242). O atendimento à condição de estacionariedade da planta é dado mediante a superação de um percentual mínimo de sinais aceitos como estacionários, sendo este percentual também uma escolha do usuário.

Tomando como exemplo a unidade de destilação que emprega o AspenPlus, o conjunto dos sinais escolhidos para uso no SSD consiste de 28 variáveis de processo assinaladas como tal no historiador de dados, sendo 10 sinais de vazão e 18 de

temperatura. Deve-se ressaltar, contudo, que a real dimensão do conjunto **ee** pode variar ao longo do tempo em função de mudanças nos critérios de escolha ao longo da operação.

O método de detecção escolhido pela operação da planta é o assim chamado método *estatístico* (Equação 353). Para fins de compatibilização com as análises anteriores neste texto e em virtude do emprego dos dados tabelados por Young [79] para o teste de hipóteses, os dados de processo serão testados com a estística Cs (226), cujos resultados são análogos à estatística R (Equações 224,353).

Na Tabela 11 estão apresentados os percentuais de pontos considerados estacionários para um universo de 8 sinais (6 de vazão: 'Fn'; 2 de temperatura: 'Tn'). Os dados das variáveis de processo referentes a cada *tag* e a indicação de estacionariedade foram obtidos do historiador de dados ao longo de um período de 23 dias consecutivos. Como pode ser notado nos valores apresentados, os pontos efetivamente assinalados como estacionários pelo RTO o são em uma proporção muito superior do que o previsto pelo teste de hipóteses do método usado.

Tabela 11 – Percentual dos pontos considerados estacionários para 8 diferentes variáveis. $rEE|_{Cs}$: percentual suposto pelo teste de hipóteses de Cs; $rEE|_{RTO}$: percentual efetivamente identificado pelo RTO como estacionário

Tag	$rEE\|_{Cs}$	$rEE\|_{RTO}$
F1	0	98.3
F2	3.2	81.1
T1	0	97.3
T2	0	90.9
F3	0	97.1
F4	4.8	99.5
F5	0	90.1
F6	6.8	84.4

Da análise da Tabela 11 fica evidente o uso em larga escala de valores de tolerância (T na Equação (353)) que se superpõem aos valores calculados da variância por diferenças sucessivas, s_d^2. Outro modo equivalente de obter o efeito destas tolerâncias consiste em executar-se o teste de hipóteses com os valores de s_d e s calculados como o previsto porém manipulando-se os valores críticos de aceitação. A

Figura 60 mostra que, dada a conveniente manipulação destes valores críticos, de forma equivalente ao uso de tolerâncias para s_d, pode-se conseguir níveis de detecção de estacionariedade similares aos observados na Tabela 11.

Contudo, se a tolerância for usada sem um apropriado ajuste no tamanho da janela de dados de análise, a detecção pode ficar muito tendenciosa para o lado da estacionariedade, aumentando a inércia das transições de detecção e causando um atraso nestas mudanças de estado. Este fato é exemplificado na Figura 61, onde é apresentado um trecho normalizado do sinal F2. Neste caso, em que ocorrem sucessivas mudanças de patamar no sinal, o efeito líquido é o de que a não estacionariedade seja indicada com atraso, correspondendo justamente aos momentos em que o sinal já retornou a um estado de variabilidade reduzida.

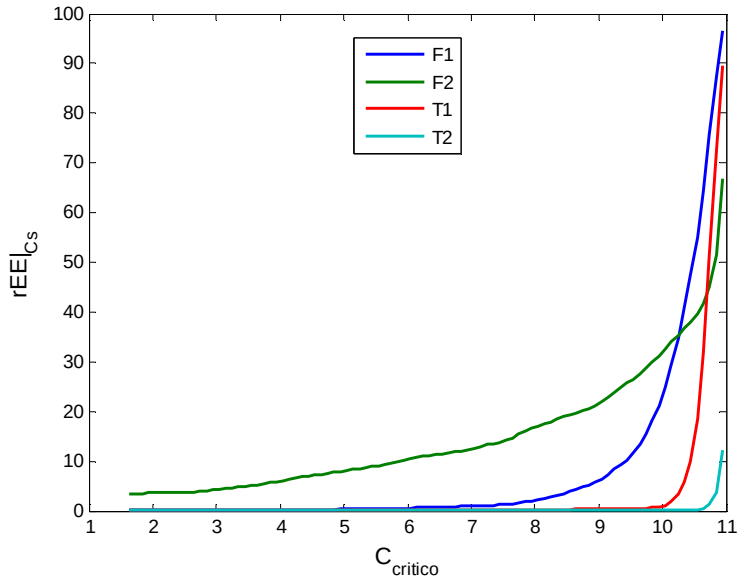


Figura 60 - Percentual dos pontos considerados estacionários em função da manipulação do valor de $C_{critico}$ usado no teste de hipóteses para quatro sinais de variáveis de processo. O valor de $C_{critico}$ calculado de acordo com as premissas do método é igual a 1,64.

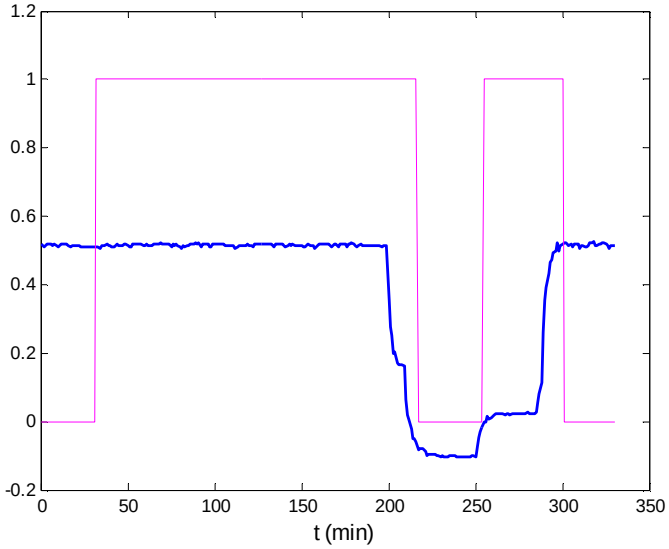


Figura 61 – Trecho normalizado de um sinal de vazão(azul) e indicação de estacionariedade do RTO (rosa), onde 1=estacionário e 0 = não estacionário

Pode parecer destoante o baixo nível de indicação de estacionariedade produzido pelo teste de hipóteses de C_s (ou R) na ausência de manipulações de tolerâncias e valores críticos, como visto na Tabela 11. Contudo, vale a pena lembrar que tal fato já foi observado na Seção 3.2.3.1 e, entre outros fatores, está relacionado a circunstâncias alheias ao processo, como a frequência de amostragem ou aos filtros de condicionamento aplicados aos sinais.

Quando aplicado este teste ao trecho de sinal contido na Figura 61, o resultado indica que nenhum ponto está estacionário. Isto pode parecer um pouco surpreendente pois a análise visual indica que, ao menos no trecho do início até cerca de t=200 min, o sinal aparenta ter um comportamento "calmo" o bastante para ser considerado estacionário. Porém vale lembrar que este teste é muito sensível à variabilidade de curtíssimo prazo, sendo afetado pelo pré-processamento do sinal no SDCD e pelo período de amostragem. Se observarmos o que ocorre nestes primeiros 200 minutos em uma escala gráfica mais apropriada (vide Figura 62) pode-se notar que, independentemente da amplitude da variação, o que importa é que há um padrão de autocorrelação no sinal que faz com que a variação entre pontos sucessivos seja menor do que o esperado caso não houvesse dependência e o valor de cada ponto fosse fruto apenas de um processo puramente aleatório. Como consequência, o valor de s_d^2 na Equação (225) torna-se pequeno em relação à variância em relação à média, como pode

ser acompanhado pela Figura 63, que mostra a evolução dos termos s_d^2 e $2s^2$ ao longo dos primeiros 200 minutos do trecho. Como consequência, tanto C como seu análogo normalizado, C_s, têm seus valores aumentados para além do valor crítico de aceitação de estacionariedade.

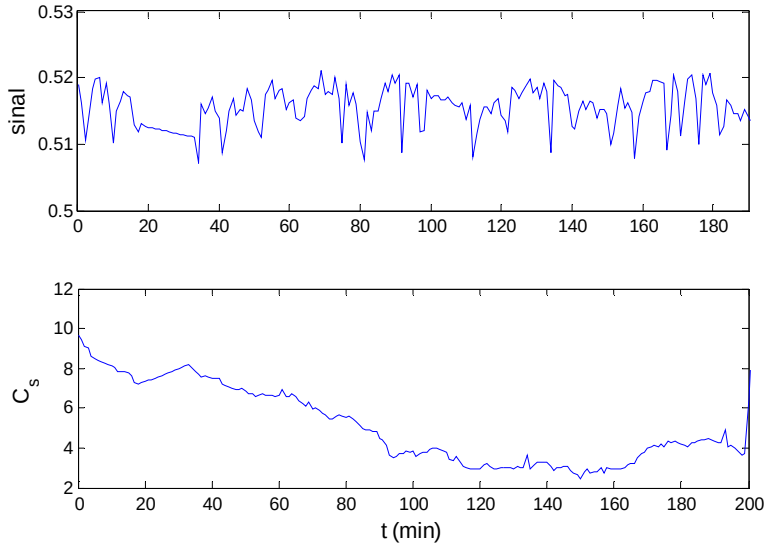


Figura 62 – Acima: Detalhe do sinal apresentado na Figura 61. Abaixo: valor da estatística Cs neste trecho. Valor crítico de acordo com as premissas do método é igual a 1,64.

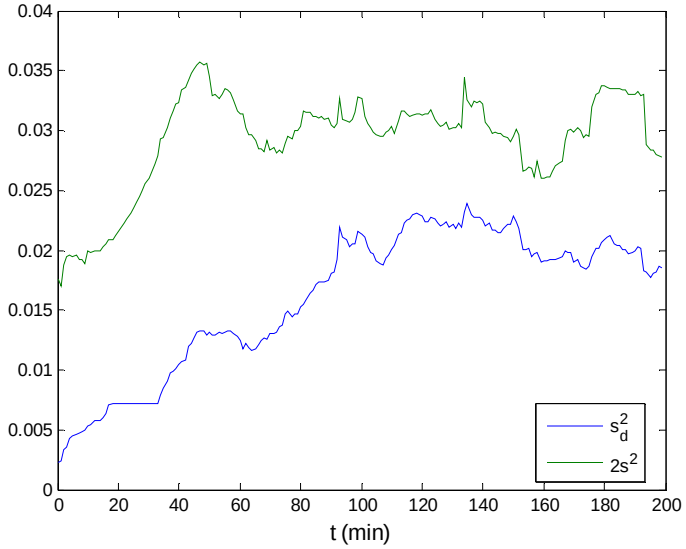


Figura 63 – Evolução de s_d^2 e $2s^2$ ao longo dos primeiros 200 minutos do sinal apresentado na Figura 61.

4.2. Adaptação e Otimização

Os sistemas de RTO usados neste texto como exemplos de implementações fazem uso de uma janela de dados cujas informações sobre as variáveis medidas são obtidas dos registros do historiador de dados da planta. Segundo a convenção adotada no presente trabalho, tal janela é representada por $\left[\mathbf{Za(obj)}\right]_{j,N,Tam}$, sendo que para aplicações de destilação em refino de petróleo comumente é adotada uma janela de uma hora de duração com intervalo de amostragem tipicamente na faixa T_{am}= 0,5 a 1 min, o que implica N = 120 a 60, respectivamente. A função objetivo para ambos os sistemas consiste no somatório dos desvios quadráticos (planta,modelo) ponderados pelas variâncias de cada medição, como apresentado na Equação (357).

$$\begin{aligned}
&\boldsymbol{\Theta}_j(\mathbf{upd}) = \underset{\boldsymbol{\Theta}_j(\mathbf{upd})}{\arg\min}\left(\underbrace{\sum_{i=1}^{\dim(\mathbf{obj})} \frac{1}{s_i^2}\left(\overline{\mathbf{Za}(\mathbf{obj}(i))} - \mathbf{Zm}(\mathbf{obj}(i))\right)^2}_{S} \right) \\
&s.a \\
&\begin{cases} \boldsymbol{f}_{sis} \\ \boldsymbol{gm} \end{cases} \\
&\overline{\mathbf{Za}(\mathbf{obj}(i))} = \frac{\sum_{k=j-N+1}^{j} \mathbf{Za}_k(\mathbf{obj}(i))}{N}
\end{aligned} \tag{357}$$

Deve-se notar que há uma importante diferença entre a função objetivo empregada pelos RTOs comerciais e aquela prevista pelo método da máxima verossimilhança apresentado na Equação (272) e reduzido à forma da Equação (282) em caso de distribuição gaussiana dos erros aditivos. Por razões provavelmente relacionadas à facilidade de implementação, a função objetivo dos *softwares* de RTO reduz a janela de dados de cada variável a um só valor, correspondente ao valor médio das medições contidas nesta janela.

Esta modificação causa mudanças que vão além da simplicidade introduzida nos cálculos, uma vez que a forma apresentada não condiz com as pretensas propriedades estatísticas esperadas do uso do método de máxima verossimilhança aplicado a problemas de estimação de parâmetros e reconciliação de dados.

Outro ponto a ser destacado é que, da forma como apresentado na Equação (357), os elementos s_i cumprem a função de espelharem o desvio-padrão das medições. Fica também evidente a suposição prévia de que os erros são independentes pois a formulação é equivalente ao emprego de uma matriz variância-covariância com termos nulos fora da diagonal principal, o que pode ser percebido da comparação com a Equação (316).

Nota-se, nas implementações industriais, que a escolha dos conjuntos **upd** e **obj** é feita de modo muito livre, baseados em procedimentos empíricos. No caso do RTO da Aspentech o número de variáveis que foram incluídas no conjunto **obj** foi de 49, considerando-se um período de análise de 3 meses (1000 rodadas). Este conjunto de dados de análise embasará as demais análises apresentadas nesta Seção.

É interessante notar que as premissas que suportam as escolhas do conjunto **obj**, apresentadas na Equação (120), nem sempre são respeitadas. Segundo estas premissas, um fator importante é que **obj** indique variáveis medidas, que são os valores não-nulos de **Za**, e que correspondem aos valores observados sob o efeito direto da corrupção do sinal experimental na Equação (119). Contudo, o que se observa no uso prático de tais sistemas é a inclusão de variáveis estimadas (**est**), as quais não são medidas, em **obj**. Um caso típico de ocorrência diz respeito a parâmetros de caracterização da carga, que pertencem a **est** e são incluídos frequentemente na função objetivo do problema de adaptação. Neste caso, na falta do valor observado **Za**(**est**) são usados valores fixos arbitrariamente escolhidos que, na prática, funcionam como limitadores de variação ao redor do valor fixo usado. Embora haja um apelo intuitivo nesta abordagem, ela não é prevista no contexto do método da máxima verossimilhança e induz a ocorrência de bias nas variáveis estimadas.

Não é comum, nem na literatura técnica nem nos sistemas comerciais de RTO, a apresentação de ferramentas de diagnóstico do sistema em uso contínuo. No decorrer desta Seção será analisada a informação produzida pela atuação de um RTO comercial usado em uma unidade de destilação de petróleo cujo modelo é da ordem de grandeza de 1.10^5 Equações. Embora esta análise não tenha pretensões de ser exaustiva nas conclusões, a intenção é apresentar alguns fatos característicos destas implementações de larga escala, baseadas no RTO em duas camadas e que ficam ocultas sob a grande quantidade de informações gerada por tais sistemas.

Assim sendo, pode-se tomar como ponto de partida a influência de cada variável incluída em **obj** sobre o valor da função objetivo (S na Equação 357). Sob a validade de

todos os requisitos R1-R6 (pág. 161) e com a correta configuração dos valores das variâncias dos erros de cada medida, é esperado que as influências normalizadas (358) de cada componente de **obj** sejam similares entre si.

Para o período estudado, que consistiu de 1000 rodadas consecutivas, o RTO obteve convergência de seu algoritmo de estimação/reconciliação em 59,7% das vezes. O intervalo de tempo entre duas rodadas *convergidas* consecutivas do RTO é tal que o 10°, 50° e 90° percentis de sua distribuição são, respectivamente, de 0.8, 1.28 e 4.86 horas. Como observado na Figura 64, em que são apresentados os valores das contribuições normalizadas (358), é comum que algumas poucas variáveis em **obj** sejam responsáveis pela maior parte do valor da função objetivo em (357).

Na Tabela 12 são apresentadas detalhadamente as contribuições das 10 variáveis que apresentam os maiores valores de 50° percentil (mediana), juntamente com uma descrição (tag) que serve para que se conheça a natureza da variável medida (F, vazão, T, temperatura). Além da comparação do valor mediano das contribuições também é de se ressaltar a sua grande variação, como visto nos valores do 10° e 90° percentis. Como exemplo, as duas variáveis ranqueadas no topo da Tabela 2 variam cerca de 50 pontos percentuais entre estes percentis. Na verdade, alguma discrepância entre a contribuição das diversas variáveis em **obj** para a função objetivo é esperada na medida em que as variáveis são expressas em unidades diferentes e tal efeito não seja plenamente compensado pelas respectivas variâncias. Contudo, este fato não explica a elevada variação desta contribuição ao longo da operação, que pode ser justificada pelo não atendimento dos requisitos R1-R6 face a diferentes cenários de operação, ainda que não seja possível discriminar quais requisitos não são antendidos.

$$ct_k = \frac{\frac{1}{s_k^2}\left(\overline{\mathbf{Za}(\mathbf{obj}(k))} - \mathbf{Zm}(\mathbf{obj}(k))\right)^2}{\sum_{i=1}^{\dim(\mathbf{obj})} \frac{1}{s_i^2}\left(\overline{\mathbf{Za}(\mathbf{obj}(i))} - \mathbf{Zm}(\mathbf{obj}(i))\right)^2} \qquad (358)$$

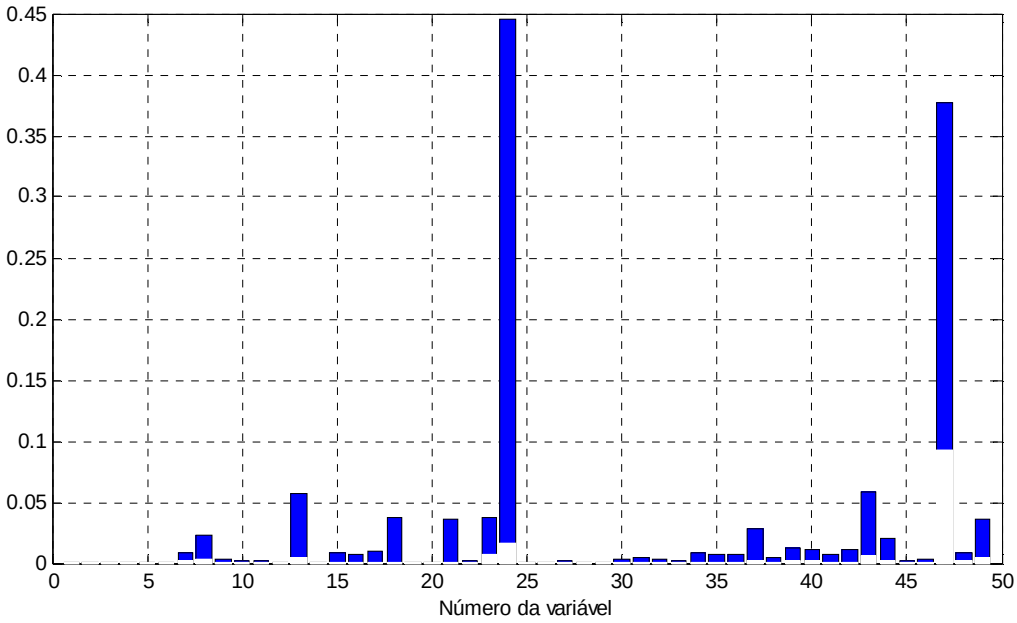


Figura 64 – Intervalo 1°-3° quartis da contribuição normalizada de cada variável em **obj** para a função objetivo do problema de adaptação. O número da variável indicado refere-se apenas à posição relativa no vetor **obj**.

Tabela 12 – Valores dos percentis das contribuições normalizadas (%) das dez variáveis que mais influenciam o valor da função objetivo de adaptação. As variáveis estão ordenadas em termos do 50° percentil (P50).

Ordem Contribuição P50	Tag	∈ ms?	P50	P90	P10
1	T01	S	21.53	54.83	1.15
2	T02	S	15.99	54.33	0.19
3	F01	S	2.37	11.39	0.10
4	T03	S	1.95	7.49	0.48
5	T04	S	1.68	14.21	0.35
6	T05	S	1.41	7.01	2.05E-06
7	T06	S	1.17	5.94	0.04
8	F07	S	1.12	7.39	4.54E-03
9	KNOT(8)	**N**	0.96	4.08	0.09
10	T07	S	0.65	2.17	0.12

O diagnóstico da operação de um RTO real não é simples pois não se tem acesso a todas as informações que impactam e que descrevem o comportamento real da planta. Por este motivo, especial atenção é dada, nesta análise, à variabilidade apresentada pelos resultados, na medida em que pode-se comparar a frequência com que as perturbações mais importantes ocorrem (mudanças de carga) com a variabilidade apresentada pelos

resultados produzidos pelo RTO, tais como valores das funções-objetivo, variáveis adaptadas, manipuladas e expectativa de desempenho econômico.

Neste contexto, vale a pena verificar o comportamento das variáveis estimadas pelo RTO que dizem respeito à qualidade da carga processada na unidade. Em refinarias não é usual a caracterização molecular do petróleo processado. Ao invés disso, são usadas análises de propriedades físico-químicas globais e de frações (cortes) de petróleo. Uma análise de emprego disseminado é a relação entre faixas de destilação e seus respectivos rendimentos volumétricos, resultando em curvas de destilação de ponto de ebulição verdadeiro (PEV), ASTM-D86 e correlatas, que diferem pelo método empregado para a destilação. Na operação convencional de uma refinaria não se dispõem de informação *online* das curvas de destilação da carga que está sendo processada, embora se disponha de bancos de dados de análises dos petróleos empregados. Estas análises podem diferir dos valores reais devido a diversos fatores:

- o banco de dados pode apresentar análises defasadas, dado que a qualidade do petróleo em cada poço modifica-se ao longo da história da sua exploração.

- pode haver mudanças de composição do petróleo devido à logística de armazenamento e distribuição desde o poço até o tanque final de uso. Isto costuma causar a perda de componentes mais voláteis.

- as regras de mistura das propriedades dos diversos petróleos que compõem a carga podem não representar perfeitamente a curva de destilação resultante.

- eventualmente, correntes internas da refinaria são reprocessadas, sendo misturadas ao *blend* de petróleo para compor a carga da unidade.

Por conta destes fatores, e uma vez que a caracterização da carga tem um grande impacto na determinação da qualidade e da quantidade dos produtos gerados, é comum incluir, dentre as variáveis estimadas, alguns parâmetros (*knots*) relacionados ao ajuste da inclinação da curva de destilação (temperatura de destilação X volume destilado acumulado). Esta curva é dividida em seções, sendo que cada *knot* representa um fator multiplicativo da inclinação da curva de destilação em cada seção. No presente caso, 11 *knots* são graus de liberdade do problema de adaptação do modelo, e serviriam para, com

base nas informações das demais variáveis medidas e com base no modelo de funcionamento da coluna de destilação, estimar a qualidade da carga que está sendo processada.

A variabilidade das estimativas destes *knots* pode ser acompanhada na Figura 65, onde é mostrada a distribuição dos valores estimados de cada *knot* (note-se que a carga de referencia corresponde ao valor de 1 para os *knots*). Pode-se também acompanhar, na mesma Figura, a diferença relativa entre duas estimativas consecutivas de cada *knot*. Note-se que, além de uma grande variabilidade total, a variação relativa entre duas estimativas consecutivas de cada knot possui amplitude muito elevada, incompatível com aquela encontrada no petróleo real. Enquanto a carga de petróleo nesta unidade muda aproximadamente uma vez por semana, com três trocas de tanque contidas neste intervalo, variações relativas de mais de 20% são frequentes entre duas estimativas consecutivas (intervalo de cerca de 1,5h), para muitos knots. Além disto, como os *knots* são estimados de forma independente, esta variação pode se distribuir em sentidos opostos entre *knots* contíguos, possivelmente dando origem a curvas de destilação que, não façam sentido físico.

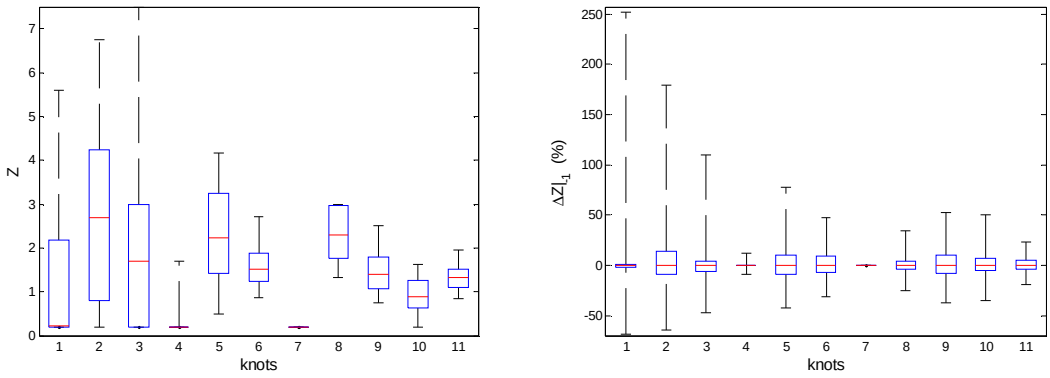


Figura 65 – Esquerda: Distribuição dos valores estimados dos knots. Direita: Distribuição do desvio relativo entre duas estimativas consecutivas dos knots.

Na Tabela 13 são mostradas as restrições impostas aos valores estimados, que fazem parte das inequações ***gm*** em (357), assim como o percentual de vezes em que as estimativas ativaram alguma restrição. Isto quer dizer que a variabilidade só não é maior devido à limitação da restrição imposta, que pretensamente impõe a realidade física às estimativas. Contudo, este procedimento é ilusório. Não será esta limitação que imporá a realidade fenomenológica, pois o fato de o valor estimado não poder ultrapassar certo limite apenas assegura que o usuário não se depare com valores insólitos, mas não

agrega confiabilidade nem significância física ao resultado. De fato, a variação que o algoritmo de otimização imporia à variável presa em uma restrição será acomodada em outros parâmetros estimados com o fito de minimizar a função objetivo. No fim das contas, o que se consegue é descontar os custos da baixa estimabilidade de um parâmetro no bias introduzido na estimação de todos os demais parâmetros estimados. Casos de estimativas pobres aprisionadas em limites pretensamente físicos introduzidos no problema de adaptação também são reportados em Quelhas, Jesus e Pinto [50].

Tabela 13 – Restrições impostas aos valores estimados dos *knots* e percentual das rodadas convergidas em que os valores estimados recaíram sobre alguma restrição.

Knot	Lim_{inf}	Lim_{sup}	% rest. ativa
1	0.2	10	42.5
2	0.2	10	13.4
3	0.2	7.5	31.5
4	0.2	7.5	81.7
5	0.2	7.5	0.7
6	0.2	7.5	0.2
7	0.2	7.5	92.1
8	0.2	3	19.4
9	0.2	3	1.2
10	0.2	3	14.6
11	0.2	3	0.3

A elevada variabilidade também se expressa nos valores da função objetivo do problema de adaptação, S na Equação (357), e pode ser notada na Figura 66, onde é possível observar a diferença relativa de S entre dois ciclos consecutivos convergidos, ΔS (Equação 359). Na ausência de mudanças físicas no processo ocorridas no curto prazo (entre dois ciclos do RTO) é difícil justificar esta magnitude de variação em termos de ocorrências reais. A única explicação cabível para a variabilidade da qualidade global do ajuste são a múltipla variação dos requisitos *R1-R6* e também a habilidade do método numérico de otimização em encontrar soluções correspondentes a extremos globais da função objetivo [50].

$$\Delta S = 100\,\frac{S_{j+1} - S_j}{S_j} \tag{359}$$

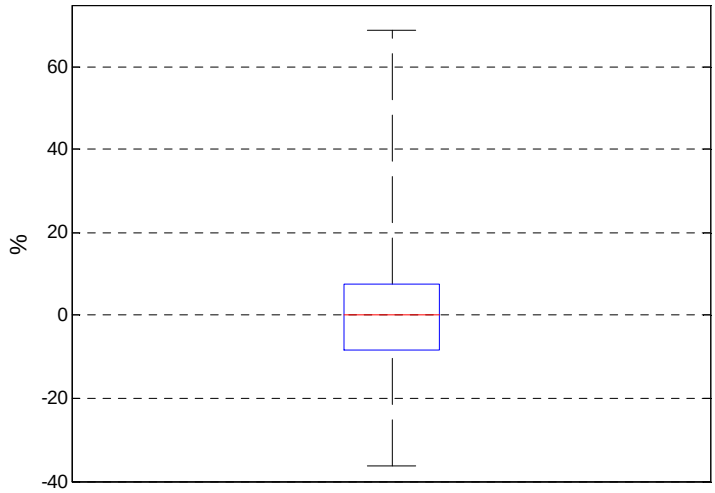


Figura 66 – Distribuição de ΔS, diferença relativa do valor da função objetivo de adaptação, entre dois ciclos consecutivos convergidos.

A Figura 67 mostra a grande diferença entre duas formas de verificar a eficiência do procedimento de otimização em tempo real. Enquanto Δ_{prev} (360) indica a variação relativa do lucro sob a ótica do otimizador do RTO, ou seja, comparando os valores calculados do lucro antes e depois da otimização dentro de um mesmo ciclo, Δ_{verif} (361), indica o ganho relativo do lucro comparando o valor encontrado no início de um ciclo com sua a indicação de lucro dada pelo RTO ao fim do ciclo anterior. Note-se que, embora ambas as métricas padeçam do fato de se referirem ao lucro visto pelos olhos do modelo do RTO, o uso de Δ_{prev} é uma medida com viés muito otimista, pois sempre reporta valores não negativos, enquanto Δ_{verif} incorpora os efeitos sofridos pela informação após passar pelo filtro representado pelo processo.

$$\Delta_{prev} = 100 \frac{Lm_{j+1}\big|_j - Lm_j\big|_j}{Lm_j\big|_j} \tag{360}$$

$$\Delta_{verif} = 100 \frac{Lm_j\big|_j - Lm_j\big|_{j-1}}{Lm_j\big|_{j-1}} \tag{361}$$

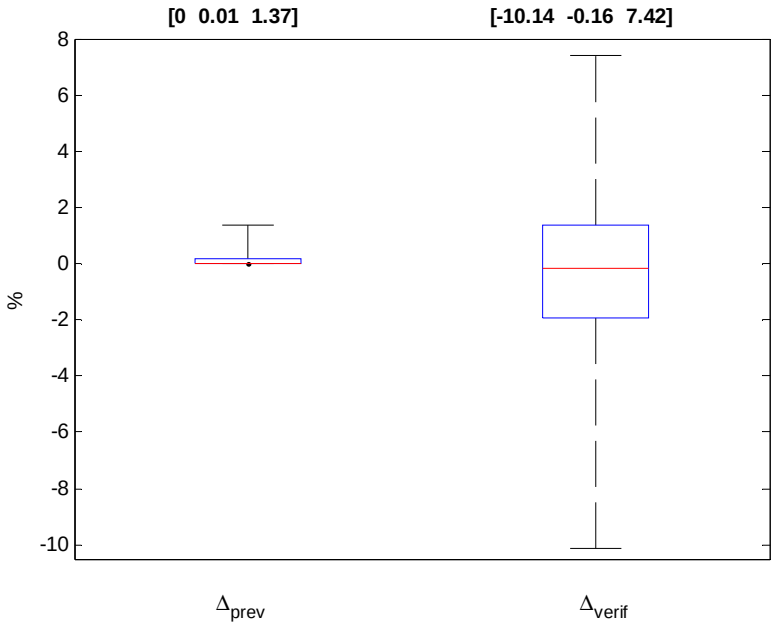


Figura 67 – Distribuição dos valores de Δ_{prev} e de Δ_{verif} para o lucro do RTO. No topo do gráfico estão indicados os precentis [P5 P50 P95] de cada distribuição.

5. Estudo de Caso

Neste capítulo será apresentado um estudo de implantação e execução de um sistema de otimização em tempo real com o objetivo de exemplificar alguns dos conceitos discutidos nos capítulos anteriores. De modo a facilitar o diálogo com a literatura técnica da área, foi escolhido o modelo reacional de Williams-Otto [113], que comumente é usado em estudos de RTO [25, 26, 44,45,4,48,49].

5.1. Apresentação do Problema

O diagrama esquemático do processo é mostrado na Figura 68. O núcleo deste processo consiste nas reações descritas na Equação (362), que ocorrem em um reator dotado de um sistema de controle capaz de manter a temperatura reacional no valor T_R. Este reator é alimentado pelas correntes F_A e F_B, que contém, respectivamente, os componentes puros A e B, assim como pela corrente de reciclo do processo, F_{rec}.

A corrente efluente do reator passa por um súbito resfriamento para estancar o prosseguimento das reações. Em seguida, todo o componente G é separado por decantação e removido na corrente F_G.

Por meio de um processo de separação, parte do componente P é removida da corrente F_E gerando o produto F_P. A corrente residual da separação é particionada entre a purga F_D e a corrente de reciclo que volta ao reator em função da razão de reciclo α.

$$\begin{cases} r1:\ A + B \rightarrow C \\ r2: C + B \rightarrow P + E \\ r3: P + C \rightarrow G \end{cases} \qquad (362)$$

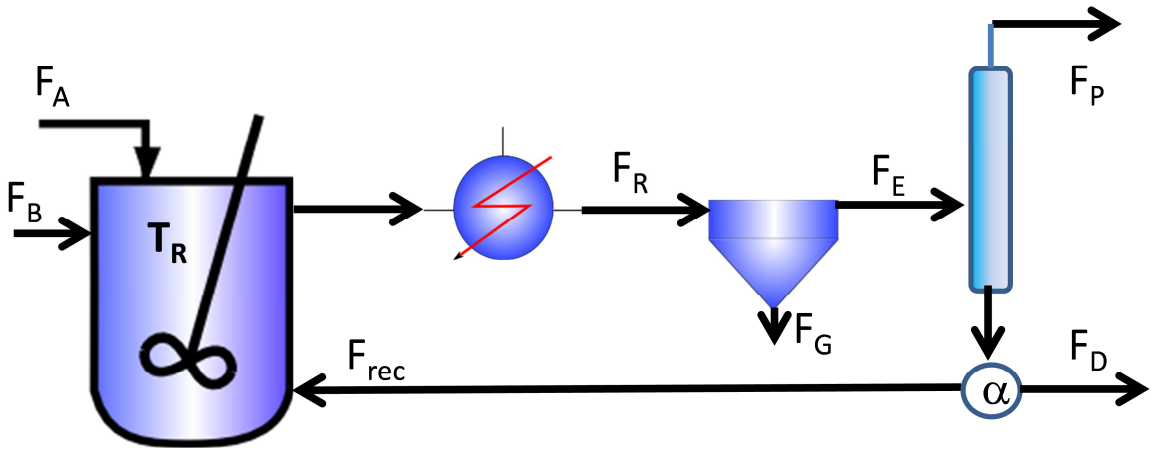


Figura 68 – Diagrama esquemático do processo de Williams-Otto.

Os balanços mássicos do processo apresentado na Figura 68 são descritos por meio das Equações (363-388), nas quais as vazões $F_{\bullet}$, bem como as frações $w_{\bullet}$, são representadas em base mássica. Uma modificação em relação à representação original diz respeito ao processo de separação do produto P, representado na Equação (388), onde é apresentada a relação entre a fração de recuperação f_P e a vazão de alimentação do separador, F_E.

$$F_A + F_B + F_{rec} - F_R = 0 \tag{363}$$

$$F_{rec} + \frac{\alpha}{(\alpha - 1)}(F_A + F_b - F_p) = 0 \tag{364}$$

$$F_R - F_E - F_G = 0 \tag{365}$$

$$F_D + F_P - F_E + F_{rec} = 0 \tag{366}$$

$$k_i = A_i e^{(-B_i / Tr)}, \quad i = 1..3 \tag{367}$$

$$F_A - F_R.w_a - k_1 V_R w_a w_b + F_{rec} w_{a,Frec} = 0 \tag{368}$$

$$F_B - F_R.w_b - k_1 V_R w_a w_b - k_2 V_R w_b w_c + F_{rec} w_{b,Frec} = 0 \tag{369}$$

$$2k_1V_Rw_aw_b - F_R.w_c - 2k_2V_Rw_bw_c - k_3V_Rw_cw_p + F_{rec}w_{c,Frec} = 0 \quad (370)$$

$$2k_2V_Rw_bw_c - F_Rw_e + F_{rec}w_{e,Frec} = 0 \quad (371)$$

$$1.5k_3V_Rw_cw_p - F_Rw_g = 0 \quad (372)$$

$$k_2V_Rw_bw_c - F_Rw_p - 0.5k_3V_Rw_cw_p + F_{rec}w_{p,Frec} = 0 \quad (373)$$

$$F_P - F_Ew_{p,Fe}f_P = 0 \quad (374)$$

$$F_G - F_Rw_{g,Fr} = 0 \quad (375)$$

$$w_{a,Fe} - w_a/(1-w_g) = 0 \quad (376)$$

$$w_{b,Fe} - w_b/(1-w_g) = 0 \quad (377)$$

$$w_{c,Fe} - w_c/(1-w_g) = 0 \quad (378)$$

$$w_{e,Fe} - w_e/(1-w_g) = 0 \quad (379)$$

$$w_{p,Fe} - w_p/(1-w_g) = 0 \quad (380)$$

$$w_{g,Fe} = 0 \quad (381)$$

$$w_{a,Frec} - F_Ew_{a,Fe}/(F_e - F_p) = 0 \quad (382)$$

$$w_{b,Frec} - F_Ew_{b,Fe}/(F_e - F_p) = 0 \quad (383)$$

$$w_{c,Frec} - F_E w_{c,Fe}/(F_e - F_p) = 0 \tag{384}$$

$$w_{e,Frec} - F_E w_{e,Fe}/(F_e - F_p) = 0 \tag{385}$$

$$w_{p,Frec} - (F_E w_{p,Fe} - F_P)/(F_e - F_p) = 0 \tag{386}$$

$$w_{g,Frec} = 0 \tag{387}$$

$$\begin{cases} F_E \le 9.10^4 \ \ kg/h: \\ \qquad f_P - \left(0{,}7783(10^{-5}F_E)^3 - 0{,}9276(10^{-5}F_E)^2 - 0{,}328(10^{-5}F_E) + 0{,}83\right) = 0 \\ F_E > 9.10^4 \ \ kg/h: \\ \qquad f_P = 0{,}35 \end{cases} \tag{388}$$

Embora o conjunto **Z** contenha todas as variáveis descritas nas Equações (363-388), será adotado, para fins de simplicidade, o conjunto mostrado na Equação (389), cujos valores típicos, particionados entre variáveis necessárias, **Z**(**in**) e consequentes, **Z**(**out**), está apresentado na Tabela 14.

$$\{Z\} = \left\{F_A, F_B, T_R, \alpha, A_{1..3}, B_{1..3}, V_R, w_a, w_b, w_c, w_p, w_e, w_g, F_P, F_{rec}, F_E, F_D, F_G\right\} \tag{389}$$

Tabela 14 – Valores típicos das variáveis contidas em **Z**.

Z(in)		
F_A	6577,1	kg/h
F_B	15127	kg/h
T_R	355,37	K
α	0,5275	-
A_1	5,98E+09	1/h
A_2	2,60E+12	1/h
A_3	9,63E+15	1/h
B_1	6666,7	K
B_2	8333,3	K
B_3	11111,0	K
Vr	2104,7	kg

Z(out)		
wa	0,121	-
wb	0,385	-
wc	0,025	-
wp	0,084	-
we	0,346	-
wg	0,038	-
F_P	2160,5	kg/h
F_{rec}	21828,0	kg/h
F_E	41863	kg/h
F_D	17880	kg/h
F_G	1669	kg/h

O produto P é aquele de maior valor de venda e constitui o foco de operação da unidade. As correntes de saída do decantador, F_G, e de purga, F_D, possuem um valor secundário, como visto na Equação (390).

Os custos da unidade são relacionados à compra de matéria prima (F_A e F_B) e às utilidades associadas ao processamento da carga da unidade de separação, F_E, conforme expresso na Equação (391), de modo que o lucro advindo da operação deste processo é descrito pela Equação (392). O valor do lucro correspondente aos valores contidos na Tabela 14 é de \$920,0/h.

$$receita = 0{,}3F_P + 0{,}0068(F_D + F_G) \quad \$/h \tag{390}$$

$$custo = (0{,}02F_A + 0{,}03F_B) + 15(F_E / 41818)^2 \quad \$/h \tag{391}$$

$$L = 5{,}13(receita - custo) \quad \$/h \tag{392}$$

5.2. Estruturas Possíveis do RTO e Cenários de Operação

De modo a configurar o RTO no presente estudo de caso, serão consideradas diversas estruturas possíveis, no espírito das considerações apresentadas no Capítulo 2. Note-se que a exploração de todas as possibilidades de estrutura contidas nas premissas anteriormente descritas resulta em uma quantidade explosiva de estruturas a serem consideradas, o que fez com que alguns cuidados tenham sido tomados em nome da garantia da exequibilidade computacional do estudo:

- o conjunto de variáveis necessárias, **in**, é fixado *a priori* (vide Tabela 14), assim como o conjunto de variáveis medidas, **ms** (vide Tabela 15), ou seja, supõe-se que não há liberdade de escolha da instrumentação.

- as escolhas dos conjuntos de variáveis atualizáveis, **upd**, e de variáveis incluídas na função objetivo de adaptação, **obj**, são feitas sobre subconjuntos de **Z (**vide Tabela 15) de modo a limitar o conjunto total de alternativas.

- são impostas as seguintes limitações aos conjuntos que definem a estrutura do RTO: $2 \le \dim(\mathbf{upd}) \le 3$, $3 \le \dim(\mathbf{obj}) \le 5$

À semelhança da Equação (125), o universo de escolhas referente às possíveis estruturas é dado pela Equação (393). De acordo com as condições propostas na presente seção e obedecendo às regras de escolhas formuladas nas Equações (110, 111 e 121) resulta um total de 7228 estruturas possíveis, conforme apresentado na Equação (394).

$$\mathbf{Rto}(x) = \left\{ \mathbf{Gupd}(y), \mathbf{Gobj}(z) \,|_{\mathbf{Gupd}(y)} \right\} = \left\{ \mathbf{upd}, \mathbf{obj} \right\}_{z|y} \tag{393}$$

$$\dim\left(\mathbf{Rto}\right) = 7228 \tag{394}$$

Tabela 15 – Resumo das propriedades e dos limites admissíveis das variáveis em **Z**

Nome	Medido?	Adapta?	Objetivo Adaptação?	Mínimo	Máximo	Unidade
F_A	1	1	0	1814	11340	kg/h
F_B	1	0	0	9072	21319	kg/h
T_R	1	0	0	335	380	K
α	1	0	0	0	0,9	
A_1	0	1	0	1,00E+09	1,00E+10	1/h
A_2	0	1	0	5,00E+11	1,00E+13	1/h
A_3	0	1	0	1,00E+15	8,00E+16	1/h
B_1	0	1	0	2000	10000	K
B_2	0	1	0	4000	15000	K
B_3	0	1	0	5000	20000	K
V_R	0	1	0	907	2722	kg
w_a	1	1	1	0	1	
w_b	1	1	1	0	1	
w_c	1	1	1	0	1	
w_p	1	1	1	0	1	
w_e	1	0	1	0	1	
w_g	1	0	0	0	1	
F_P	1	1	0	227	6804	kg/h
F_{rec}	1	1	0	0	90718	kg/h

As estruturas contidas no conjunto **Rto** só podem ser discriminadas dentro de um contexto que leve em conta todas as influências sobre o desempenho global. Por conta disto, é necessário levar em conta a natureza dos sinais medidos assim como os cenários

de variabilidade das condições impostas pelas variáveis necessárias ao longo dos diversos e sucessivos ciclos de operação do sistema.

Para o presente caso, a estrutura genérica da corrupção dos sinais medidos (Equação 395) supõe a existência de erros autocorrelacionados cuja parcela estocástica pode apresentar média não-nula, caracterizando a existência de *bias* de medição. Os valores numéricos assumidos para os parâmetros na Equação (395) são apresentados na Tabela 16, sendo possível notar que esta Tabela, na verdade, resume três diferentes configurações de erros cujas parametrizações distinguem-se pelas combinações assumidas para o *bias* das variáveis F_A, F_P e F_{rec}, conforme explicitado na Tabela 17.

$$\varepsilon(n) = \phi\varepsilon(n-1) + \delta(n); \quad \delta \sim N(bias, \sigma) \tag{395}$$

Tabela 16 – Parametrização dos erros de medição. σ e *bias* indicam valores relativos em relação ao valor real da variável.

Nomes	**σ (%)**	***ϕ***	***bias* (%)**
F_A	1	0.4	[0,1,-1]
F_B	0	0	0
T_R	0	0	0
α	0	0	0
w_a	1	0.4	0
w_b	1	0.4	0
w_c	1	0.4	0
w_p	1	0.4	0
w_e	1	0.4	0
w_g	1	0.4	0
F_P	1	0.4	[0,1,1]
F_{rec}	1	0.4	[0,1,1]

Tabela 17 – ConFigurações de bias (%) na instrumentação

	F_A	F_P	F_{rec}
instr1	0	0	0
instr2	1	1	1
instr3	-1	1	1

É assumido que o processo pode enfrentar cinco situações típicas de operação (cenários), cada qual constando de seis ciclos sucessivos de diferentes condições impostas pelas variáveis necessárias, **Z**(**in**). Os diversos cenários são apresentados nas Tabelas 19-23. O primeiro cenário, C1, apresenta condições de contorno constantes ao

longo de todos os ciclos, com valores equivalentes aos apresentados na Tabela 14. Para os demais cenários, algumas condições sofrem alterações ao longo dos ciclos ou partem de valores distintos das condições-base contidas em C1. Para os cenários C2-C5, as Tabelas 20-23 apresentam na cor amarela as células em que as condições diferem do respectivo ciclo em C1.

Por fim, com o objetivo de comparar a influência de alterações na conformação do problema a ser resolvido nas escolhas de arquitetura do RTO foram estudadas, paralelamente, três versões alternativas do problema, nomeadas e descritas de acordo com a Tabela 18.

Tabela 18 – Versões do problema a ser resolvido pelo RTO

sem restr	As propriedades e limites das variáveis em **Z** são aquelas descritas na Tabela 15. Há duas variáveis de decisão, a saber, F_B e T_R. **df** = [2 3], **u** = Z(**df**).
restr	Difere do problema anterior em um ponto: há restrição ao valor máximo da fração mássica do componente E na saída do reator: $0 \leq w_e < 0{,}39$
restr3u	Possui a mesma restrição do problema *restr*. Contudo, agora o problema tem três variáveis de decisão: F_B , T_R e α. **df** = [2 3 4]

Tabela 19 – Valores consecutivos de **Z**(**in**) para o cenário de operação C1.

C1	Ciclo do RTO					
Nomes	**1**	**2**	**3**	**4**	**5**	**6**
F_A	6577	6577	6577	6577	6577	6577
F_B	15127	15127	15127	15127	15127	15127
T_R	355,4	355,4	355,4	355,4	355,4	355,4
α	0,528	0,528	0,528	0,528	0,528	0,528
A_1	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09
A_2	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12
A_3	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15
B_1	6667	6667	6667	6667	6667	6667
B_2	8333	8333	8333	8333	8333	8333
B_3	11111	11111	11111	11111	11111	11111
V_R	2104,7	2104,7	2104,7	2104,7	2104,7	2104,7

Tabela 20 – Valores consecutivos de **Z(in)** para o cenário de operação C2.

C2	Ciclo do RTO					
Nomes	**1**	**2**	**3**	**4**	**5**	**6**
F_A	6577	6577	6577	5591	5591	5591
F_B	15127	15127	15127	15127	15127	15127
T_R	355,4	355,4	355,4	355,4	355,4	355,4
α	0,528	0,528	0,528	0,528	0,528	0,528
A_1	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09
A_2	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12
A_3	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15
B_1	6667	6667	6667	6667	6667	6667
B_2	8333	8333	8333	8333	8333	8333
B_3	11111	11111	11111	11111	11111	11111
V_R	2104,7	2104,7	2104,7	2104,7	2104,7	2104,7

Tabela 21 – Valores consecutivos de **Z(in)** para o cenário de operação C3.

C3	Ciclo do RTO					
Nomes	**1**	**2**	**3**	**4**	**5**	**6**
F_A	6577	6577	6577	7564	7564	7564
F_B	15127	15127	15127	15127	15127	15127
T_R	355,4	355,4	355,4	355,4	355,4	355,4
α	0,528	0,528	0,528	0,528	0,528	0,528
A_1	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09	5,98E+09
A_2	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12
A_3	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15
B_1	6667	6667	6667	6667	6667	6667
B_2	8333	8333	8333	8333	8333	8333
B_3	11111	11111	11111	11111	11111	11111
V_R	2104,7	2104,7	2104,7	2104,7	2104,7	2104,7

Tabela 22 – Valores consecutivos de **Z(in)** para o cenário de operação C4.

C4	Ciclo do RTO					
Nomes	**1**	**2**	**3**	**4**	**5**	**6**
F_A	6577	6577	6577	7564	7564	7564
F_B	15127	15127	15127	15127	15127	15127
T_R	355,4	355,4	355,4	355,4	355,4	355,4
α	0,528	0,528	0,528	0,528	0,528	0,528
A_1	5,98E+09	5,98E+09	5,98E+09	5,38E+09	5,38E+09	5,38E+09
A_2	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12	2,60E+12
A_3	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15	9,63E+15
B_1	6667	6667	6667	6333	6333	6333
B_2	8333	8333	8333	8333	8333	8333
B_3	11111	11111	11111	11111	11667	11667
V_R	2104,7	2104,7	2104,7	2104,7	2104,7	2104,7

Tabela 23 – Valores consecutivos de **Z**(**in**) para o cenário de operação C5.

C5	Ciclo do RTO					
Nomes	**1**	**2**	**3**	**4**	**5**	**6**
F_A	6577	6577	6577	6577	6577	6577
F_B	15127	15127	15127	15127	15127	15127
T_R	355,4	355,4	355,4	355,4	355,4	355,4
α	0,528	0,528	0,528	0,528	0,528	0,528
A_1	6,27E+09	6,27E+09	6,27E+09	6,27E+09	6,27E+09	6,27E+09
A_2	2,47E+12	2,47E+12	2,47E+12	2,47E+12	2,47E+12	2,47E+12
A_3	1,01E+16	1,01E+16	1,01E+16	1,01E+16	1,01E+16	1,01E+16
B_1	7000	7000	7000	7000	7000	7000
B_2	7917	7917	7917	7917	7917	7917
B_3	10555	10555	10555	10555	10555	10555
V_R	2104,7	2104,7	2104,7	2104,7	2104,7	2104,7

5.3. Escolha da Estrutura do RTO

De forma a sintetizar o problema colocado nas seções anteriores, formulando-o de acordo com a simbologia proposta neste trabalho, a implantação deste sistema de RTO tem como premissas a representação completa (Equação 396) das variáveis pertinentes descritas em (Equação 389) e a equivalência da representação disponível do processo com a realidade (Equações 397,398).

Deste modo, dada a malha de instrumentação previamente existente (Equação 399), existem 7228 alternativas de formulação dos conjuntos de variáveis atualizáveis e de variáveis participantes da função objetivo do problema de adaptação de acordo com as condições propostas pelas Equações (Equações 400-401).

$$\{\mathbf{Zm}\} = \{\mathbf{Z}\} \tag{396}$$

$$\mathcal{P}m = \mathcal{P} \tag{397}$$

$$Tproc \Rightarrow \left\{ \left(\mathbf{Zm(in)} \xrightarrow{eqs.(351\text{-}376)} \mathbf{Zm(out)} \right) \Leftrightarrow \left(\mathbf{Z(in)} \xrightarrow{eqs.(351\text{-}376)} \mathbf{Z(out)} \right) \right\} \tag{398}$$

$$\mathbf{ms} = [1,2,3,4,12,13,14,15,16,17,18,19] \tag{399}$$

$$\mathbf{upd} \subset [1,5,6,7,8,9,10,11,12,13,14,15,18,19], \quad 2 \leq \dim(\mathbf{upd}) \leq 3 \tag{400}$$

$$\mathbf{obj} \subseteq [12,13,14,15,16], \quad 3 \leq \dim(\mathbf{obj}) \leq 5 \tag{401}$$

O sistema de RTO será avaliado em função das respostas produzidas ao longo de cinco cenários (Equação 402) constituídos por seis ciclos de atuação consecutivos (Equação 403). Levando-se em conta que são supostas três diferentes configurações de corrupção dos sinais medidos, o conjunto completo de condições operacionais, *CO*, é constituído de 15 instâncias (Equação 404).

$$\mathbf{Zcen} \in \{C1, C2, C3, C4, C5\} \tag{402}$$

$$\mathbf{Zcen} = [\mathbf{Z}_0 \ldots \mathbf{Z}_{numc-1}], \; numc = 6 \tag{403}$$

$$CO = \{C1, C2, C3, C4, C5\} \; X \; \{instr1, instr2, instr3\}; \; \dim(CO) = NC = 15 \tag{404}$$

Colocado o problema desta forma, a escolha de uma estrutura em detrimento de outra deve ocorrer em função de um modo de discriminar as vantagens associadas a esta escolha. Cada uma das 7228 estruturas deve ter a ela associada uma métrica, genericamente expressa por Ω na Equação (Equação 405), que convenientemente caracterize o desempenho ao longo de cada ciclo de atuação consecutivo, em todas as condições operacionais e de forma replicada, de modo a reproduzir a característica estocástica da informação medida

$$\Omega|_x = \Omega(i, j, k, \mathbf{Rto}(x))\Big|_{i=CO(1)..CO(NC),\; j=1..numc,\; k=1..NR} \tag{405}$$

A natureza aberta do problema genérico de implantação do RTO não permite a definição apriorística de uma expressão única e definitiva de Ω, aplicável a quaisquer processos. Serão as especificidades contidas em cada implementação, as diretrizes operacionais, de segurança e financeiras que ditarão o modo mais conveniente de medir as vantagens associadas a cada configuração.

Ainda que ciente da natureza particular e subjetiva das motivação implícitas na formulação desta métrica, pretende-se propor, no presente trabalho, uma medida que possua, conceitualmente, apelo prático para uma grande variedade de processos químicos e que permita avançar com a decisão da estrutura do RTO. Com este pensamento em mente, esta formulação assume como premissa o fato de que há três pontos principais aos quais deve ser dada atenção durante a operação da planta sob ação de um RTO. Estes pontos podem ser descritos e quantificados da seguinte forma:

1) **Sub-optimalidade do objetivo financeiro**: Uma vez que toda a construção do método de RTO em duas camadas culmina na maximização do lucro operacional, é razoável valorar o afastamento dos resultados obtidos em relação àquele que potencialmente seria obtido caso os procedimentos garantissem desempenho ideal. Neste espírito, a formulação do vetor $\mathbf{\Delta L_x}$ (Equação 406) contém os afastamentos relativos a este desempenho ideal, *Lo*, da *x*-ésima estrutura analisada, ao longo de *numc* ciclos consecutivos, de cada uma das NC condições operacionais, replicadas NR vezes.

$$\mathbf{\Delta L}_x = \left. \frac{100\left(L_{j+1}(i,k)\big|_j - Lo_{j+1}(i,k)\big|_j\right)}{Lo_{j+1}(i,k)\big|_j} \right|_{\mathbf{Rto}(x)}, \quad i = 1..NC,\ k = 1:NR,\ j = 1:numc \tag{406}$$

Falta ainda explicar a natureza do lucro ideal, Lo. Neste ponto, vale a pena recapitular algumas informações contidas na Figura 4, na página 62. Note-se que, em cada ciclo de execução do RTO, existem quatro informações referentes ao lucro sendo produzidas: L_j, o lucro real que o processo produz no ciclo j; Lm_j, a imagem de L_j vista com os olhos do RTO; $Lm_{j+1}\big|_j$, o lucro que o RTO enxerga ser possível alcançar no próximo ciclo, fruto de sua atuação agora, e $L_{j+1}\big|_j$ o lucro a ser efetivamente alcançado no próximo ciclo caso as implementações sejam feitas e nenhuma variação de cenário ocorra. Além destas, podemos supor a existência de uma quinta entidade nesta lista, de natureza abstrata, originada de um ciclo de otimização paralelo a este descrito na Figura 4, fruto de uma otimização não sujeita a quaisquer dos obstáculos descritos na Seção 2.2 nem a imperfeições dos métodos de otimização numérica. Este ciclo pode ser visto na Figura 69 e esclarece o conceito contido em Lo_j.

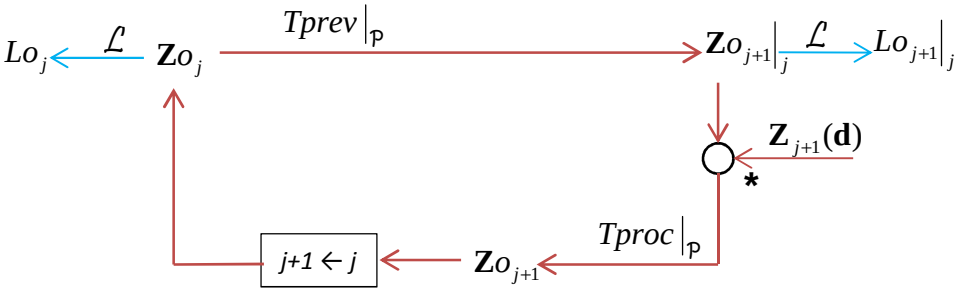


Figura 69 – Funcionamento em malha fechada de RTO baseado em modelo e otimização perfeitos

A transformação do vetor $\mathbf{\Delta L_x}$ em uma métrica de natureza escalar deve, de alguma forma, levar em conta a informação parametrizada da pdf de $\mathbf{\Delta L_x}$**.** Existem múltiplas possibilidades de realizar esta operação, a mais comum sendo uma combinação linear da média e do desvio-padrão [4, 30] da distribuição. Contudo, o uso do desvio-padrão não é de grande conveniência em distribuições assimétricas, como as esperadas nos casos presentemente estudados. Além disto, para o problema de interesse neste texto, a variabilidade embutida no conceito de desvio-padrão só importa se se manifestar no sentido de aumento do desvio da métrica estudada, não no de sua diminuição. Por este motivo, é mais interessante levar um conta informações diretamente tomadas da pdf, que independam de hipóteses preliminares sobre a simetria de sua conformação e que realcem a variabilidade no sentido das maiores consequências para o problema. A medida $\mathcal{ML}$, dada pela média ponderada da mediana e do 25° percentil, como expresso em (407), é uma alternativa para compactar a informação expressa pela pdf de $\mathbf{\Delta L_x}$, incorporando elementos da tendência central e da dispersão manifestada no sentido de interesse. A versão normalizada de $\mathcal{ML}_x$ é dada por mL_x na Equação (408).

$$\mathcal{ML}_x = \frac{\left(2P_{50}(\mathbf{\Delta L}_x)+P_{25}(\mathbf{\Delta L}_x)\right)}{3} \tag{407}$$

$$mL_x = \frac{-\mathcal{ML}_x}{\min(\boldsymbol{\mathcal{ML}})} \tag{408}$$

2) **Variabilidade das variáveis de decisão**: Embora o desempenho do resultado financeiro apareça com mais evidência entre as características a serem valorizadas, não

se deve menosprezar os efeitos indesejáveis de excessivas manipulações nos valores das variáveis de decisão. Dada a natureza estacionária da otimização, são ignoradas as consequências reais nas trajetórias das variáveis de processo, tais como a criação de sucessivos e longos transientes entre os ciclos e a imposição de um regime de operação caracterizado pela permanente perturbação das condições de contorno.

A variável $\mathbf{\Delta u_x}$ (Equação 409) procura capturar o efeito da variabilidade desnecessária associada à existência do RTO, expressando de forma relativa a variação entre ciclos das variáveis de decisão, descontada a variabilidade necessária ($\mathbf{uo}_j$ - $\mathbf{uo}_{j-1}$), referida a intervenções que idealmente deveriam ser realizadas no processo. A variável **uo** refere-se ao subconjunto de **Zo** (vide Figura 69), dado por **Zo**(**df**).

Seguindo raciocínio análogo ao anteriormente exposto para o lucro, $\mathcal{M}u_x$ (Equação 410) é a métrica que consolida a pdf da distribuição de $\mathbf{\Delta u_x}$ e mu_x é sua versão normalizada.

$$\mathbf{\Delta u}_x = 100 \left\| \frac{\left(\left(\mathbf{u}_j(i,k) - \mathbf{u}_{j-1}(i,k)\right) - \left(\mathbf{uo}_j(i,k) - \mathbf{uo}_{j-1}(i,k)\right)\right)}{\mathbf{u}_{j-1}(i,k)} \right\|_{\mathbf{Rto}(x)}, \qquad (409)$$

$$i = 1..NC,\ k = 1:NR,\ j = 2:numc$$

$$\mathcal{M}u_x = \frac{\left(2P_{50}(\mathbf{\Delta u}_x) + P_{75}(\mathbf{\Delta u}_x)\right)}{3} \qquad (410)$$

$$mu_x = \frac{\mathcal{M}u_x}{\min(\boldsymbol{\mathcal{M}u})} \qquad (411)$$

3) **Potencial de violação de restrições**: Na medida em que o RTO for incapaz de seguir o fluxo idealizado da Figura 69 perdem-se as garantias de que a previsão $\mathbf{Zm}_{j+1}\big|_j$ não produza efetivamente um vetor $\mathbf{Z}_{j+1}$ que viole restrições assumidas na conformação original do problema. Estas violações podem trazer consequências importantes tanto em termos de lucro quanto em termos de segurança operacional e não podem ser negligenciadas no projeto do RTO.

A quantificação do risco associado a estas violações tem como ponto de partida a variável $\mathbf{V}_x$ (Equações 412-413), associada ao valor médio de violações ao longo dos ciclos dos cenários de operação. A métrica associada ao potencial de violações para a x-

ésima estrutura é dada por MVx (Equação 414), sendo expressa de modo normalizado por mVx (Equação 415).

$$\mathbf{V}_x = \frac{\sum_{i=1}^{NC} \left(\frac{\sum_{j=2}^{tamc} v(j,i,k)}{tamc-1} \right) \Big|_{\mathbf{Rto}(x)}}{NC}, \quad k = 1:NR \tag{412}$$

$$v(j,i,k) = \begin{cases} 0, \; se \; \mathbf{g} < 0 \\ 1, \; caso \; contrário \end{cases} \tag{413}$$

$$\text{M}V_x = \frac{\left(2P_{50}(\mathbf{V}_x) + P_{75}(\mathbf{V}_x) \right)}{3} \tag{414}$$

$$mV_x = \frac{\text{M}V_x}{\min(\text{M}V)} \tag{415}$$

Se o objetivo for possuir uma medida unificada da conveniência que a escolha de dada estrutura traz ao sistema, é necessário de alguma forma agrupar as métricas que traduzem as consequências para o lucro, para a variabilidade desnecessária das variáveis de decisão e para o potencial de violações. A métrica $mluv_x$ (Equação 416) supõe a equivalência destas importâncias e produz uma versão normalizada que, embora provavelmente deva ser alterada para cada caso real considerado, produz resultados úteis para as análises neste presente trabalho. Nas Figuras 70, 71 e 72 pode ser observado o comportamento das diversas métricas para as três versões do problema apresentadas na Tabela 18. É importante observar que a sensibilidade relativa das métricas com a escolha da estrutura depende da métrica em questão e também da versão do problema. Embora todas as versões do problema sejam sensíveis a variações nas métricas de em função da estrutura escolhida, em algumas versões, como *semrestr*, esta sensibilidade pode ser muito amplificada, como visto da comparação da Figura 70 com as Figuras 71 e 72 ao observar-se a faixa de variação dos valores relativos das métricas.

$$mluv_x = \frac{\mathrm{m}L_x + \mathrm{m}u_x + \mathrm{m}V_x}{\min(\boldsymbol{mL}+\boldsymbol{mu}+\boldsymbol{mV})} \tag{416}$$

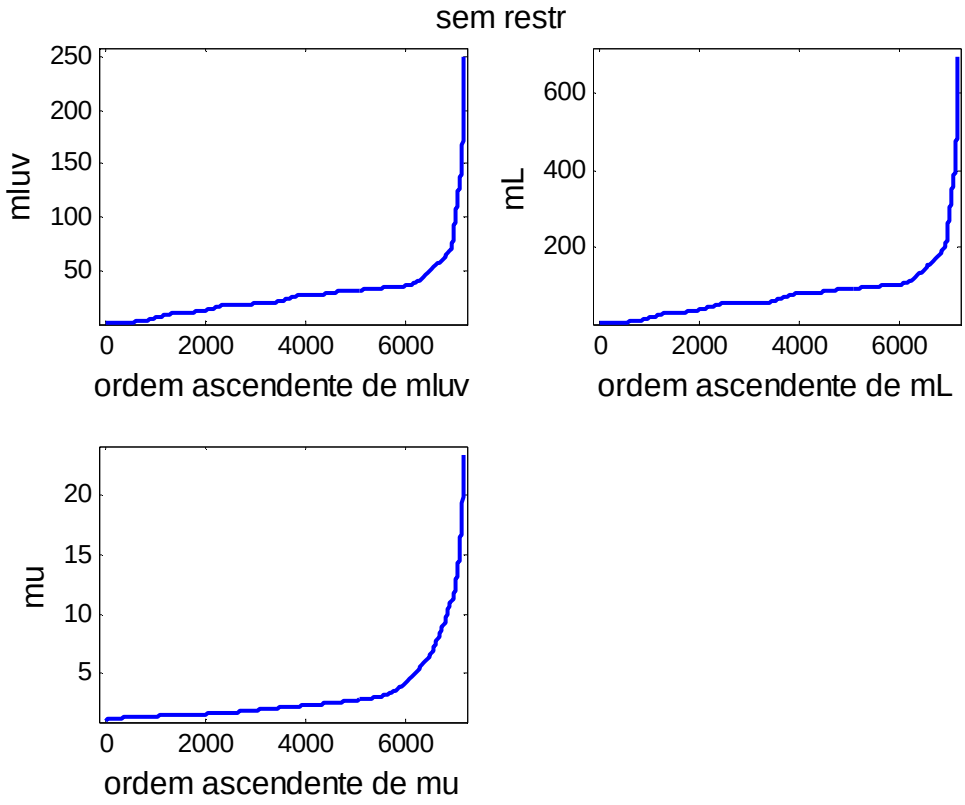


Figura 70- Valores ordenados das métricas para as estruturas disponíveis. Caso sem restrição

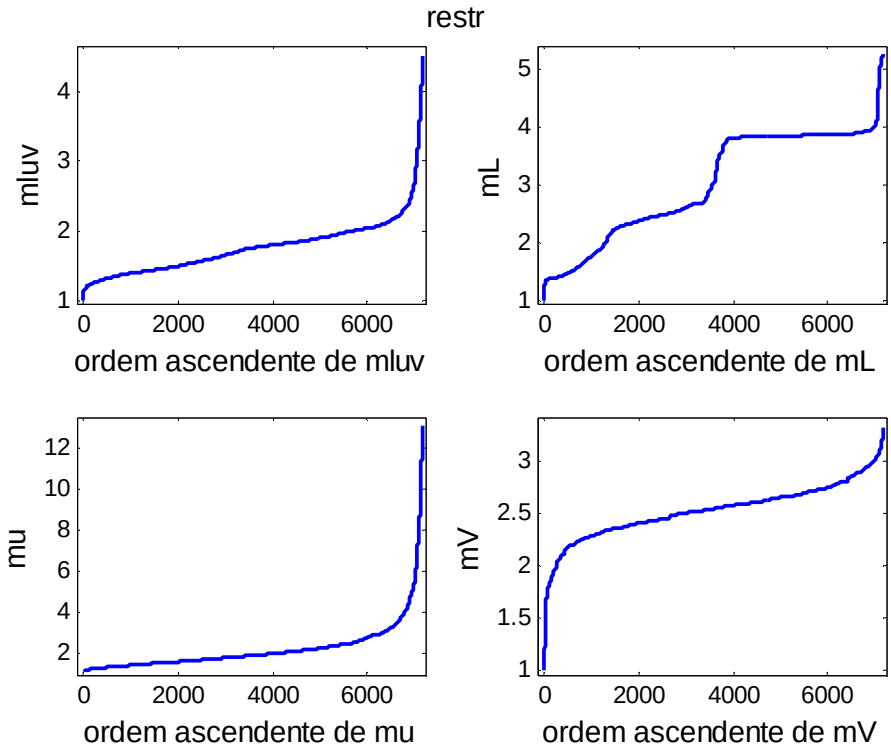


Figura 71 - Valores ordenados das métricas para as estruturas disponíveis. Caso com restrição

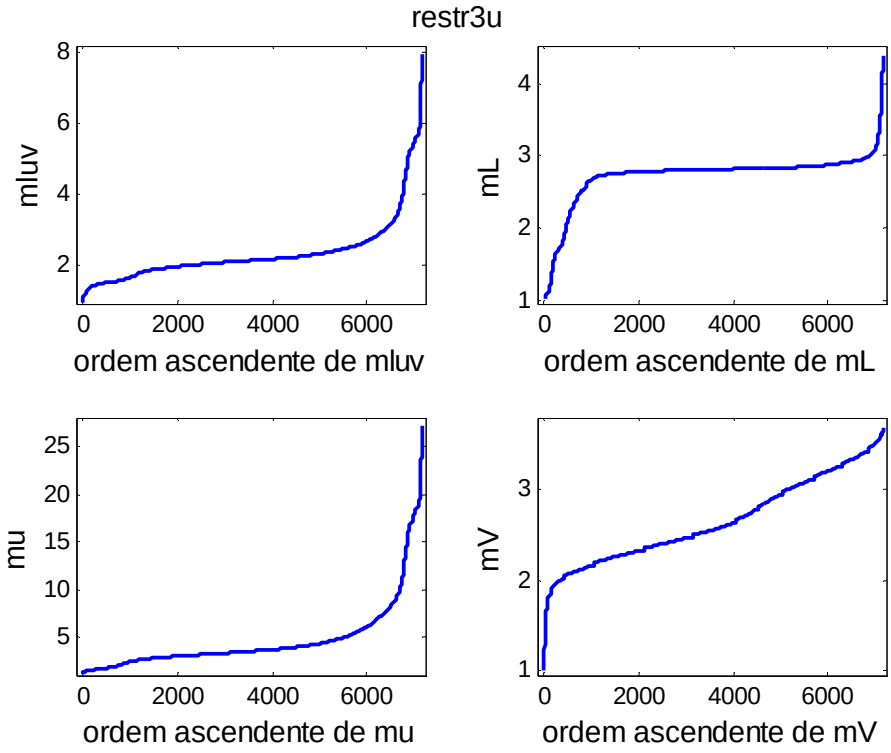


Figura 72 – Valores ordenados das métricas para as estruturas disponíveis. Caso com restrição e 3 variáveis de decisão.

É importante notar que não há qualquer garantia de que as submétricas (lucro, variáveis de decião, violações) estejam correlacionadas. Pelo contrário, como visto na Figura 73, para as três versões do problema apresentado na Tabela 18, a métrica *mluv* tem que equilibrar direções contraditórias, privilegiando soluções de compromisso que não são as melhores sob o ponto de vista de nenhuma submétrica em particular. Este fato pode ser melhor apreciado nas Tabelas 24, 25 e 26, nas quais são apresentados, para a estrutura selecionada sob o critério de dada métrica, os valores medidos pelas demais métricas. A falta de correlação entre os interesses traduzidos pelas métricas é dependente da versão do problema considerada. A Tabela 24 mostra que, para o caso *semrestr*, a melhor estrutura sob o ponto de vista da métrica agrupada *mluv* (melhor$|_{Luv}$) produz valores da métrica das variáveis de decisão, *mu*, 93% maiores do que aquele obtido com a melhor estrutura sob o ponto de vista de *mu* (melhor$|_{u}$). Por outro lado, esta estrutura melhor$|_{u}$ produz valores de *mluv* 3200% maiores que do que os valores de mluv produzidos pela estrutura melhor$|_{Luv}$.

A especificidade do problema pode conduzir a escolhas de estruturas que produzam resultados muito pobres sob o ponto de vista de algumas métricas. Tomando como exemplo os casos *restr* e *restr3u*, , se por razões operacionais ou de segurança, for mandatório que se priorize a minimização da chance de ocorrência de violações das restrições, a escolha da melhor estrutura sob o ponto de vista de mV implicaria a convivência com graus bem mais elevados, em termos relativos, de suboptimalidade do lucro e de variabilidade das variáveis de decisão, conforme mostrado nas Tabelas 25 e 26.

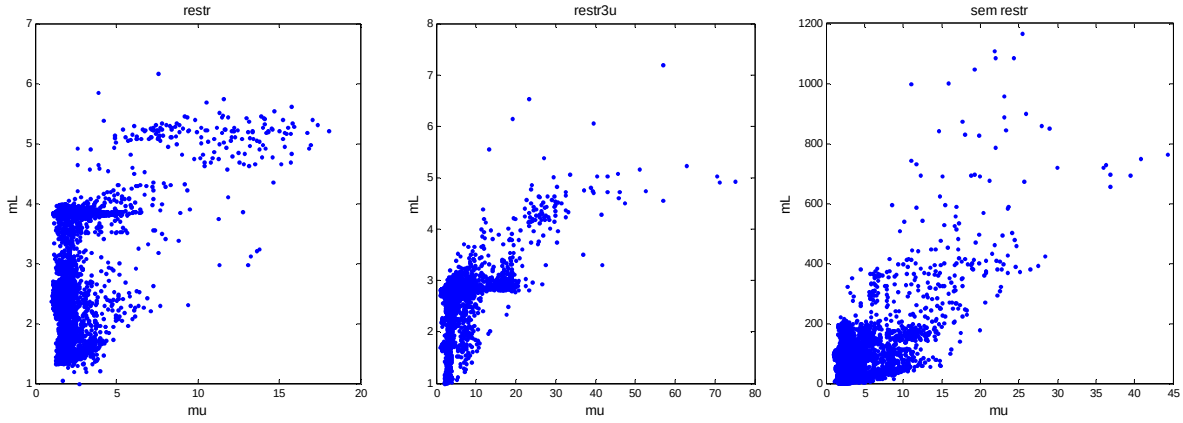


Figura 73 – Relação entre as métricas de variabilidade do lucro, *mL* e das variáveis de decisão, *mu*, para os três casos estudados.

Tabela 24 – Desempenho, sob o ponto de vista de todas as métricas, da melhor estrutura sob o ponto de vista de cada métrica

sem restr	melhor$\|_{LuV}$	melhor$\|_{L}$	melhor$\|_{u}$
mluv	1	1.00	33.33
mL	1.00	1	96.78
mu	1.93	1.93	1

Tabela 25 - Desempenho, sob o ponto de vista de todas as métricas, da melhor estrutura sob o ponto de vista de cada métrica

restr	melhor$\|_{LuV}$	melhor$\|_{L}$	melhor$\|_{u}$	melhor$\|_{V}$
mluv	1	1.25	1.42	4.04
mL	1.05	1	2.38	3.21
mu	1.70	2.69	1	13.59
mV	1.66	1.81	2.89	1

Tabela 26 - Desempenho, sob o ponto de vista de todas as métricas, da melhor estrutura sob o ponto de vista de cada métrica

restr3u	melhor$\|_{LuV}$	melhor$\|_{L}$	melhor$\|_{u}$	melhor$\|_{V}$
mluv	1	1.21	1.00	8.90
mL	1.70	1	1.70	4.37
mu	1.00	1.99	1	32.46
mV	1.55	2.14	1.55	1

Os valores dos parâmetros selecionados para **upd** e **obj** correspondentes às cinco melhores estruturas sob o critério da métrica agrupada mluv podem ser vistos na Tabela 27, assim como os parâmetros associados a diversos percentis provenientes da distribuição ordenada sob este critério podem ser vistos na Tabela 28. As cinco melhores estruturas sob a ótica de cada uma das submétricas é apresentado nas Tabelas 29, 30 e 31.

É digno de nota que as estruturas apresentadas nas Tabelas 27 a 31 corroboram o fato de que, para o problema de RTO em duas etapas, se o conjunto das informações corrompidas ou variáveis ao longo do tempo possui dimensão maior que a do conjunto de variáveis atualizáveis, dim(**crp** $\cup$ **var**) > dim(**upd**), os parâmetros atualizados não expressarão a realizada física de suas contrapartes no mundo real. Na verdade, os valores dos modificadores espelharão a projeção da variabilidade contida no espaço $\mathbb{R}^{\dim(\mathbf{crp} \cup \mathbf{var})}$

no espaço $\mathbb{R}^{\dim(\mathbf{upd})}$. Esta perda de significância física é incontornável e faz parte das limitações mais fundamentais deste tipo de abordagem, ocorrendo ainda que todas as premissas do método de máxima verossimilhança sejam atendidas e que os métodos numéricos de otimização sejam perfeitos.

A natureza puramente matemática dos valores adaptados condena uma das vantagens colaterais comumente associadas ao uso de um sistema de RTO: a de permitir o acompanhamento da variação de parâmetros imensuráveis, auxiliando na tarefa de diagnóstico e de manutenção preditiva de equipamentos e processos. Ao mesmo tempo, inibe o uso de procedimentos intuitivos de escolha dos parâmetros a serem estimados, pois o problema real a ser encarado não é o de escolher os parâmetros cuja evolução nós teríamos interesse de acompanhar, e sim escolher os parâmetros que sejam úteis, matematicamente, para conferir robustez e melhor desempenho às métricas de desempenho que norteiam os objetivos da implantação do sistema. Isto fica evidente do fato de que algumas estruturas dentre as de melhor desempenho contém variáveis que nunca se modificam sob quaisquer dos cenários, o que poderia parecer paradoxal e certamente contra-intuitivo. Tais variáveis certamente nunca seriam escolhidas por um usuário que utilizasse apenas sua experiência ou *bom senso* na análise do problema e que conhecesse os cenários de operação. Contudo, se os parâmetros forem vistos como mero instrumento de acomodação da variabilidade observada em um espaço de dimensão superior, fica mais fácil entender porque variáveis tais como a massa contida no reator, V_R, possa ser uma das variáveis atualizadas na 4ª melhor estrutura da versão *semrestr* do problema, sob a ótica da métrica agrupada mluv (Tabela 27), assim como está presente também na melhor estrutura da versão *restr3u* do problema sob a ótica das violações, mV (Tabela 31), além de também aparecer em diversas outras estruturas selecionadas como as melhores sob os outros critérios e versões do problema.

Tabela 27 – Cinco melhores estruturas sob o ponto de vista de mluv para cada versão do problema.

mluv	sem_restr	restr	restr3u
1	U(B1,B2,B3)O(wb,wc,wp,we)	U(Fa,A2,B1)O(wa,wb,wp)	U(Fa,wp,F_{rec})O(wa,wc,wp,we)
2	U(B1,B3,wc)O(wb,wp,we)	U(A2,A3,B1)O(wa,wc,wp,we)	U(wc,wp,F_{rec})O(wa,wb,wp,we)
3	U(B1,B2,B3)O(wc,wp,we)	U(A1,B2,F_{rec})O(wb,wp,we)	U(Fa,wa)O(wb,wc,we)
4	U(B1,B3,Vr)O(wc,wp,we)	U(A2,A3,B1)O(wc,wp,we)	U(Fa,wa)O(wa,wb,wc)
5	U(A3,B1,wb)O(wb,wp,we)	U(Fa,A2,B1)O(wa,wp,we)	U(Fa,F_{rec})O(wc,wp,we)

Tabela 28 – Estruturas correspondentes aos percentis indicados a partir da ordenação das estruturas pelos valores de mluv para cada versão do problema.

mluv	sem_restr	restr	restr3u
P5	U(A3,Vr,wp)O(wa,wb,wc,wp)	U(B3,Vr,F_{rec})O(wa,wb,wp)	U(Fa,wb)O(wb,wc,we)
P25	U(A2,wa,fp)O(wa,wp,we)	U(Fa,B1,wc)O(wa,wb,wp,we)	U(A3,F_{rec})O(wa,wp,we)
P50	U(A2,B3,wc)O(wa,wc,wp)	U(A2,A3,wb)O(wa,wc,wp)	U(B1,B2,wb)O(wb,wp,we)
P75	U(Fa,B2,B3)O(wa,wc,wp,we)	U(B3,wp,fp)O(wa,wc,wp,we)	U(B1,B2,fp)O(wa,wc,wp)

Tabela 29 - Cinco melhores estruturas sob o ponto de vista de mL para cada versão do problema.

mL	sem_restr	restr	restr3u
1	U(B1,B2,B3)O(wb,wc,wp,we)	U(A3,B1,wb)O(wa,wb,wc,we)	U(A3,B1,fp)O(wc,wp,we)
2	U(B1,B3,wc)O(wb,wp,we)	U(Fa,A2,B1)O(wa,wb,wp)	U(A3,B1,wb)O(wb,wc,wp,we)
3	U(B1,B2,B3)O(wc,wp,we)	U(A2,B3,Vr)O(wa,wb,wc)	U(A3,B1,wb)O(wb,wp,we)
4	U(A3,B1,wb)O(wb,wp,we)	U(A1,A3,Vr)O(wa,wb,wc,we)	U(A3,B1,wc)O(wb,wp,we)
5	U(Fa,B1,B3)O(wa,wc,wp)	U(A1,A3,wc)O(wa,wb,wc)	U(B1,B3,F_{rec})O(wc,wp,we)

Tabela 30 - Cinco melhores estruturas sob o ponto de vista de mu para cada versão do problema.

mu	sem_restr	restr	restr3u
1	U(B1,Vr,wa)O(wb,wc,we)	U(A1,wa,wc)O(wb,wc,we)	U(Fa,wp,F_{rec})O(wa,wc,wp,we)
2	U(B1,wp,fp)O(wb,wc,wp,we)	U(A1,B1,wp)O(wa,wp,we)	U(wa,wb,F_{rec})O(wb,wc,wp)
3	U(B1,wb,wc)O(wb,wp,we)	U(Fa,B1,wa)O(wb,wc,wp,we)	U(wa,fp)O(wb,wp,we)
4	U(B1,Vr,fp)O(wb,wc,we)	U(B1,Vr,wc)O(wa,wb,wp)	U(wb,fp,F_{rec})O(wa,wp,we)
5	U(A1,Vr,wp)O(wb,wc,we)	U(B1,wa,wc)O(wa,wp,we)	U(Fa,wb)O(wb,wc,wp)

Tabela 31 - Cinco melhores estruturas sob o ponto de vista de mV para cada versão do problema.

mV	sem_restr	restr	restr3u
1	U(Fa,A1)O(wa,wb,wc,wp,we)	U(A1,A3,wb)O(wb,wc,we)	U(A1,A3,Vr)O(wb,wc,we)
2	U(Fa,A1)O(wa,wb,wc,wp)	U(A3,B1,F_{rec})O(wb,wc,we)	U(A1,A3,B3)O(wb,wc,we)
3	U(Fa,A1)O(wa,wb,wc,we)	U(A3,B1)O(wb,wc,we)	U(A1,A3,fp)O(wb,wc,we)
4	U(Fa,A1)O(wa,wb,wp,we)	U(A1,A3,B1)O(wb,wc,we)	U(B1,B3)O(wb,wc,we)
5	U(Fa,A1)O(wa,wc,wp,we)	U(A1,A3)O(wb,wc,we)	U(A1,A3)O(wb,wc,we)

5.4. Desempenho do RTO

Nesta Seção serão mostrados os resultados obtidos com as estruturas de RTO escolhidas sob o critério da métrica agrupada mluv. As estruturas escolhidas correspondem à 1ª linha da Tabela 27, para cada uma das versões do problema descritos na Tabela 18.

A dimensão dos dados produzidos é consideravelmente elevada, considerando-se o número de ciclos, repetições, condições operacionais e versões do problema, de modo que é importante o uso de métricas que consolidem os resultados e subsidiem a correta interpretação dos fenômenos ocorridos. Para tanto, serão utilizadas algumas métricas sugeridas na Seção anterior assim como algumas novas serão propostas de modo a realçar aspectos relevantes dos dados.

A estrutura dos erros e da instrumentação é aquela descrita nas Tabelas 16 e 17 e os cenários são os das Tabelas 19-23, compondo um total de NC = 15 condições operacionais (vide Equação 404). Todas as as condições são replicadas (NR = 500) de modo a permitir acessar a natureza estocástica dos resultados.

Sob o ponto de vista do procedimento de atualização do modelo, são sugeridas duas métricas de desempenho. Na Equação (417) é apresentada o vetor $\Delta\boldsymbol{\theta}_{\text{m-r}}$, que está relacionado com a norma dos desvios relativos entre os valores vistos pelo RTO e a realidade para cada ciclo de cada cenário ao longo de todas as réplicas. Por outro lado, o vetor $\Delta\boldsymbol{\theta}_{\text{m(-1)}}$ (Equação 418) está relacionado com a norma da variabilidade inútil dos parâmetros atualizados entre dois ciclos consecutivos de atuação do RTO em dado cenário, por todos os cenários, ao longo de todas as réplicas.

$$\Delta\boldsymbol{\theta}_{m-r} = 100\left\| \frac{\boldsymbol{\theta}\mathrm{m_j(i,k)} - \boldsymbol{\theta}_j(i,k)}{\boldsymbol{\theta}_j(i,k)} \right\|,\ i=1..NC;\ k=1:NR;\ j=1..numc \tag{417}$$

$$\Delta\boldsymbol{\theta}_{m(-1)} = 100\left\| \frac{(\boldsymbol{\theta}\mathrm{m_j(i,k)} - \boldsymbol{\theta}m_{j-1}(i,k)) - (\boldsymbol{\theta}_j\mathrm{(i,k)} - \boldsymbol{\theta}_{j-1}(i,k))}{\boldsymbol{\theta}_{j-1}(i,k)} \right\|, \tag{418}$$

$$i=1..NC;\ k=1:NR;\ j=2..numc$$

$$\boldsymbol{\theta}m = \mathbf{Zm(upd)};\ \boldsymbol{\theta} = \mathbf{Z(upd)} \tag{419}$$

Note-se que as diferentes versões do problema amplificam de modo distinto os efeitos das imperfeições do sistema de RTO, sendo esta amplificação, colocada em ordem decrescente, dado sequencialmente pelas versões *restr3u*, *restr* e *semrestr* para os cenários C1 e C5, mostrados nas Figuras 74 a 76.

Ambos os cenários C1 e C5 contém condições constantes para todas as variáveis **in** ao longo de todos os seus ciclos. Contudo, mesmo assim os parâmetros apresentam considerável variabilidade entre ciclos consecutivos, como visto na distribuição dos valores de $\mathbf{\Delta\theta}_{m(-1)}$.

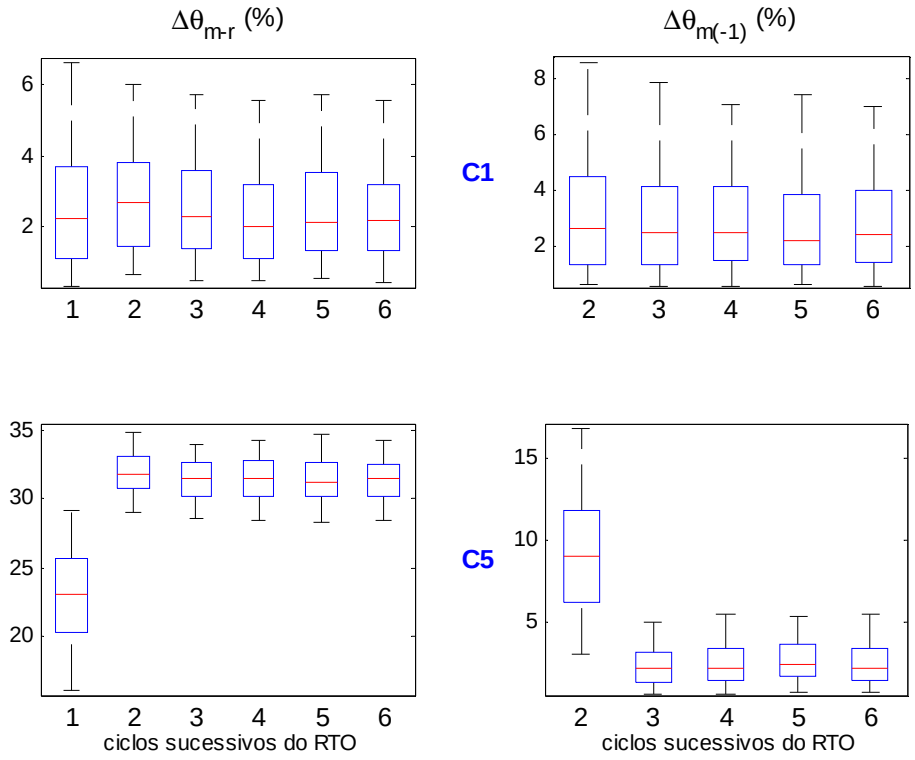


Figura 74 – Desempenho do RTO para a versão *restr* do problema sob a ótica das métricas dos parâmetros para os cenários C1 e C5. Esq: medida do afastamento dos valores reais. Dir: medida da variabilidade dos parâmetros ao longo dos ciclos do RTO.

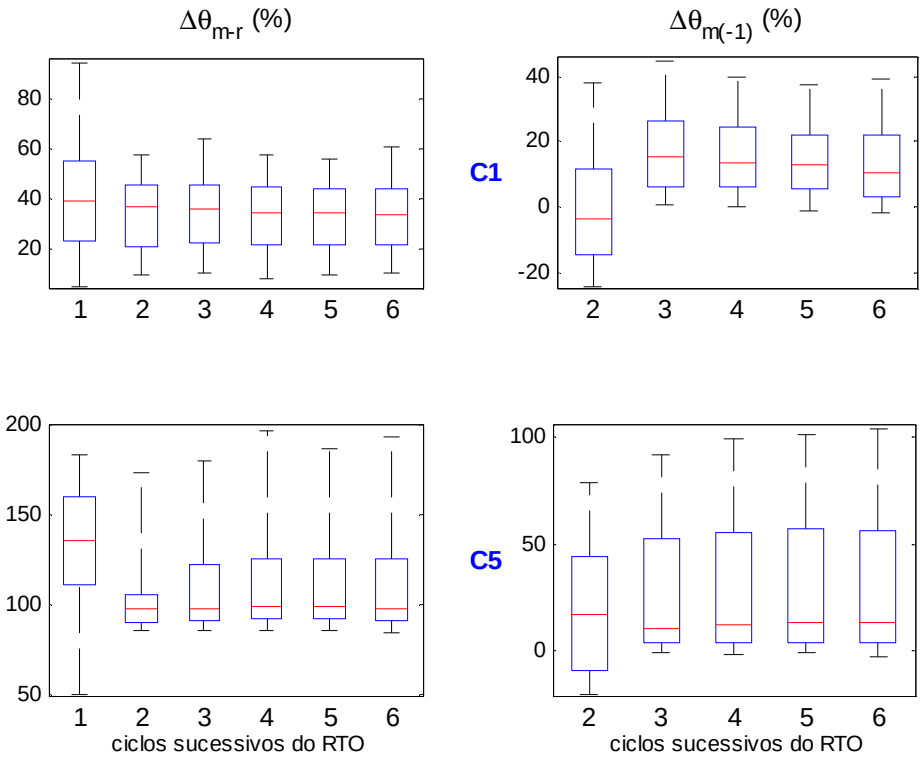


Figura 75 – Desempenho do RTO para a versão *restr3u* do problema sob a ótica das métricas dos parâmetros para os cenários C1 e C5. Esq: medida do afastamento dos valores reais. Dir: medida da variabilidade dos parâmetros ao longo dos ciclos do RTO.

Para o cenário C1, em que todas as variáveis necessárias **in** são constantes e onde não há erro de processamento da informação, tal fato mostra o efeito da projeção $\mathbb{R}^{\dim(\mathbf{crp} \cup \mathbf{var})}$ no espaço $\mathbb{R}^{\dim(\mathbf{upd})}$, referida na Seção anterior, propagado pelo RTO operando em malha fechada. Para o cenário C5, além dos efeitos contidos verificados em C1, também se conjugam os efeitos do processamento incorreto de tipo 1 (vide Seção 2.2.2) uma vez que, neste cenário, diversas variáveis de atribuição tem a elas valores incorretamente associados. Tal fato pode ser observado na Tabela 23 (página 205) que apresenta os valores das variáveis ao longo dos ciclos do cenário C5. Note-se que, apesar de constantes, as variáveis assinaladas em amarelo na Tabela 23 diferem do valor assumido *a priori* , conforme mostrado na Tabela 14, ou seja, $\mathbf{Zm}_0 \neq \mathbf{Z}_0$, configurando o processamento incorreto de tipo 1 na medida em que não há mecanismos de modificação de todos o conjunto de valores destoantes ao longo da operação.

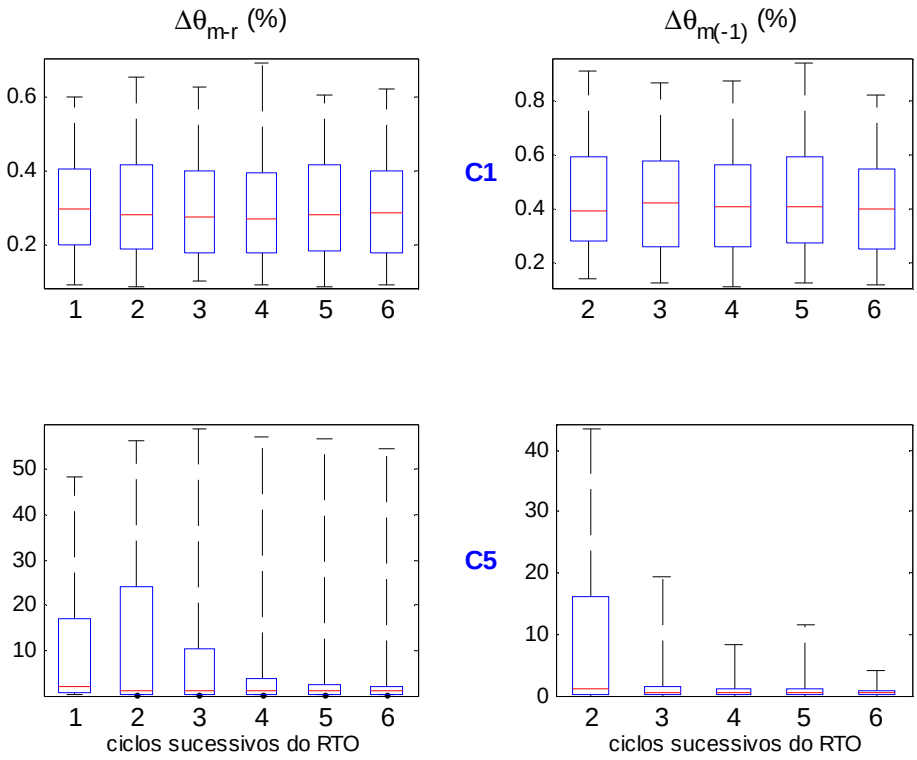


Figura 76 – Desempenho do RTO para a versão *semrestr* do problema sob a ótica das métricas dos parâmetros para os cenários C1 e C5. *Esq*: medida do afastamento dos valores reais. *Dir*: medida da variabilidade dos parâmetros ao longo dos ciclos do RTO.

Sob o ponto de vista das variáveis de decisão, a variabilidade desnecessária introduzida pelo RTO é traduzida pelas mudanças impostas a **Z**(**df**) que não estão a serviço do reposicionamento operacional devido a mudanças das condições de contorno. Esta variabilidade desnecessária das mudanças entre ciclos consecutivos de atuação, foi definida como **Δu** na Equação (409). Para o problema estudado, a distribuição de **Δu** produzida nas três versões do problema é apresentada na Figura 77. Como já obervado anteriormente, as pequenas mudanças entre as versões do problema produzem consequências apreciáveis em termos do comportamento da métrica.

A informação de variabilidade, condensada em **Δu** e apresentada na Figura 77, é a expressão da variabilidade entre duas ações consecutivas do RTO. Um panorama geral da dispersão dos valores produzidos para as variáveis de decisão pode ser visto nas Figuras 78 e 79 para a versão *restr*. Nas mesmas figuras pode-se também notar o grau de proximidade das mudanças propostas em relação ao valor correspondente ao ótimo de operação para cada cenário. Note-se que, uma vez que não há garantias de optimalidade

embutidas no procedimento do RTO em duas etapas (vide Capítulo 3) em alguns cenários (C4 e C5) as condições ótimas não estão sequer contidas na região de variabilidade esperada pela atuação do RTO.

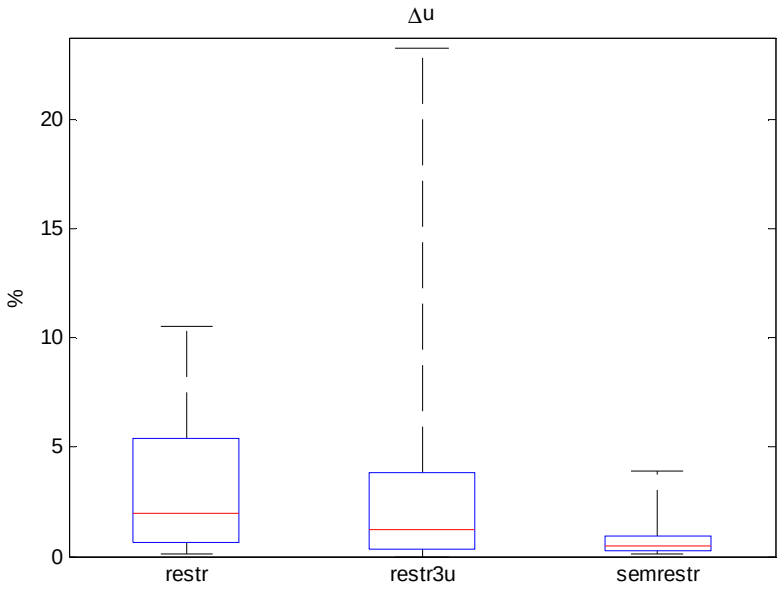


Figura 77 – Medida da variabilidade desnecessária entre ciclos consecutivos das variáveis de decisão, **Δu**, para diversas versões do problema.

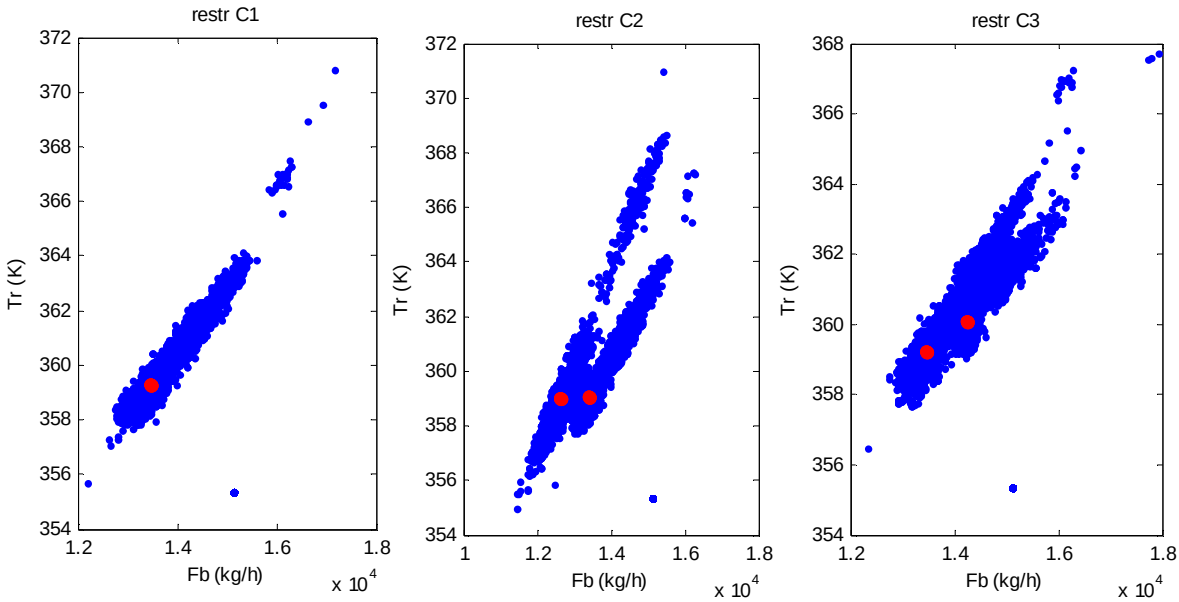


Figura 78 – Variáveis de decisão produzidas pelo RTO para o caso *restr* nos cenários C1 a C3. Em vermelho, valores que que correspondem ao lucro ótimo (a existência de mais de um valor ótimo está associada a cenários cujas condições de contorno variam ao longo dos ciclos do RTO).

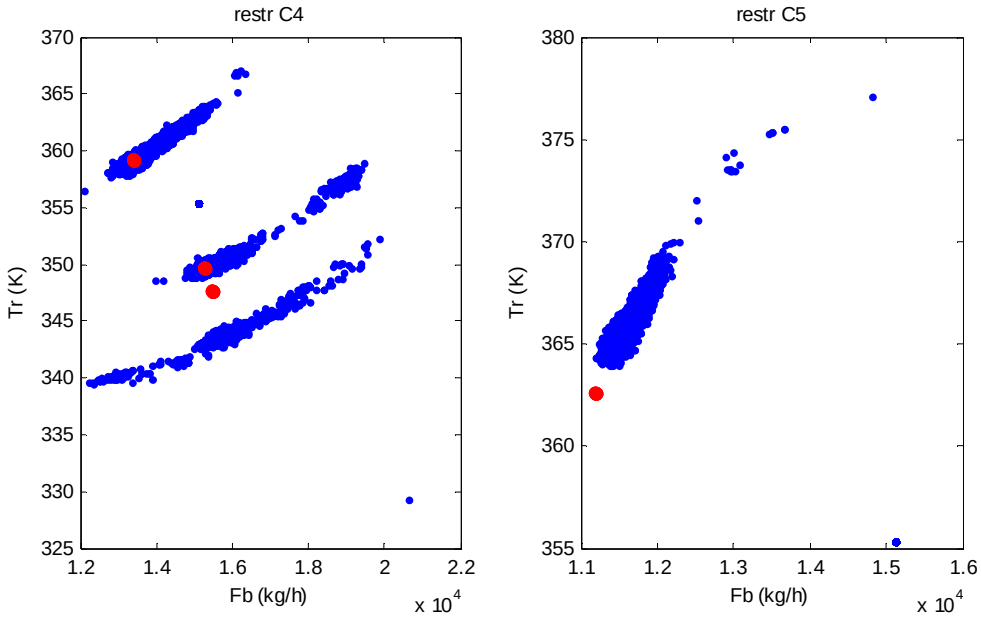


Figura 79 – Variáveis de decisão produzidas pelo RTO para o caso *restr* nos cenários C4 e C5. Em vermelho, valores que que correspondem ao lucro ótimo (a existência de mais de um valor ótimo está associada a cenários cujas condições de contorno variam ao longo dos ciclos do RTO).

Sob o ponto de vista do lucro, serão usadas nesta Seção medidas de desempenho ligeiramente diferentes da usada na Seção de seleção de estrutura. Isto porque $\mathbf{\Delta L_x}$ (Equação 406) é uma medida pontual de afastamento do ótimo, enquanto que aqui será prestada atenção ao montante de dinheiro acumulado ao longo de cada cenário. É importante frisar que o usuário só tem acesso às informações capturadas pelo RTO (**Zm**, *Lm*) e com elas deve aferir o comportamento do sistema e tomar suas decisões. Contudo, o valor efetivamente produzido, expresso em termos relativos ao acumulado sob condições ótimas, é dado por *Slucr* (Equação 420).

A distribuição dos valores de lucro relativo acumulado ao longo dos cenários pode ser observada na Figura 80, onde nota-se que a dispersão de subotimalidade real, *Slucr*, diferencia-se de forma pronunciada ao longo das versões do problema. Note-se quão sensível são os resultados são em função das pequenas mudanças introduzidas ao longo das três versões do problema. É importante também ressaltar que o problema apresentado nesta seção apresenta é relativamente bem comportado sob o ponto de vista da dependência do lucro com as variáveis de decisão. Como visto na Figura 81, a superfície é razoavelmente aplainada no centro da faixa de valores.

$$\boldsymbol{Slucr} = \frac{\sum_{j=1}^{tamc} L_{j+1}(i,k)\Big|_j}{\sum_{j=1}^{tamc} Lo_{j+1}(i,k)\Big|_j}, \quad i = 1..NC,\ k = 1:NR \tag{420}$$

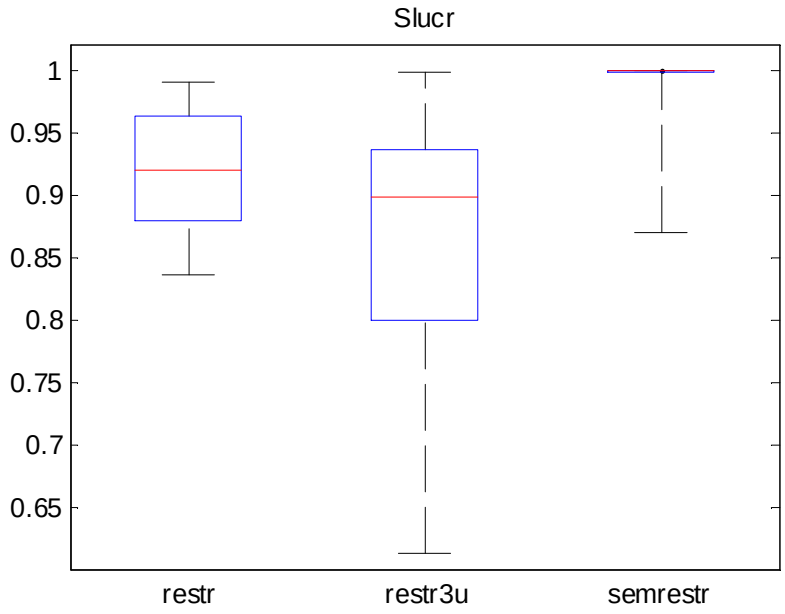


Figura 80 – Razão entre o lucro acumulado ao longo dos cenários e o lucro acumulado ótimo para diversas versões do problema.

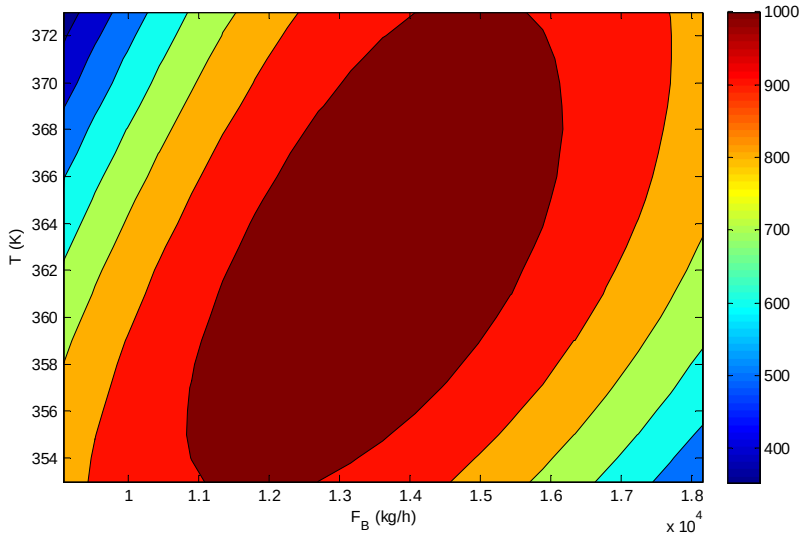


Figura 81 – Valor do lucro operacional para a condição-padrão do problema.

Sob o ponto de vista da expectativa de violação das restrições, expresso pelas Equações (412-413), os resultados esperados para as versões *restr* e *restr3u* podem ser vistos na Figura 82, onde nota-se a grande probabilidade de que cerca de um terço dos ciclos do RTO produzam sugestões de implementação que violarão as restrições do problema.

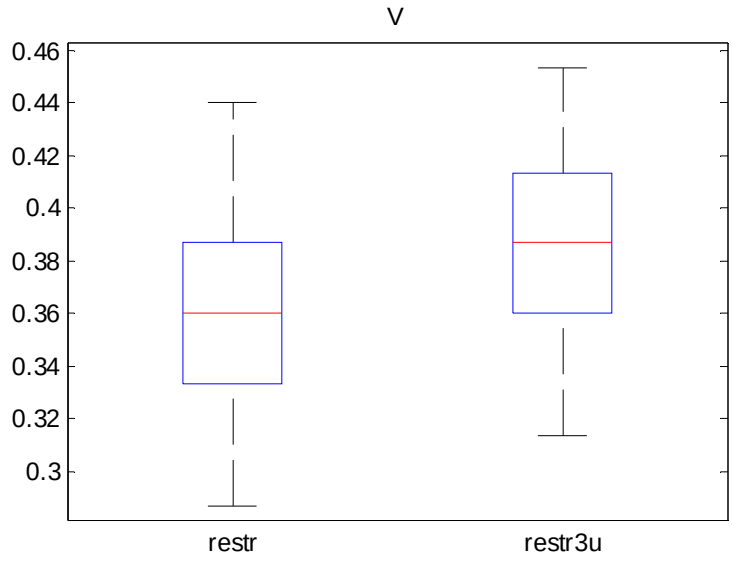


Figura 82 – Métrica de variabilidade de violação de restrições, **V**, para diversas versões do problema.

Os resultados desta Seção não são apresentados com o intuito de mostrar conclusões generalizadas sobre o comportamento de sistemas de RTO, mas sim para mostrar didaticamente o modo como se materializam muitos dos conceitos desenvolvidos nas seções anteriores a respeito das vulnerabilidades destes sistemas. Contudo , o que se pretende ressaltar é que a tarefa de seleção da estrutura do RTO é imprescindível e não pode se basear em critérios intuitivos, nem pode ser feita de forma seccionada entre os diversos componentes do RTO (SSD, adaptação, otimização...). A estrutura deve expressar a configuração integral do sistema e assim deve ser avaliada. Outro ponto importante é que a seleção da estrutura não resolve as vulnerabilidades do RTO em duas etapas: apenas o configura de modo a minimizar suas deficiências, conferindo uma robustez operacional que ele não possui em termos fundamentais. Mesmo assim, dependendo do problema em questão, mesmo a melhor estrutura pode não ser capaz de dar o desempenho esperado.

6. Discussões

Ao longo do curso deste trabalho foram apresentados diversos aspectos tanto teóricos quanto relativos às implementações industriais baseadas em *softwares* comerciais de RTO. A prática cotidiana do uso destes sistemas tem mostrado que não é fácil discernir o real ganho deste tipo de tecnologia. Normalmente, são apresentados resultados mensais de supostos ganhos devidos ao RTO. Contudo, estes valores costumam ser obtidos pela multiplicação da vazão de carga da unidade por um fator constante, expresso em $/carga da unidade. Este fator pretensamente daria conta do ganho proporcionado por sistemas de RTO, embora sua aplicabilidade não possua fundamentação conhecida. Além de proporcionar uma expectativa desconectada da realidade, este tipo de procedimento não é capaz de distinguir entre uma implementação boa e uma ruim: todas estão fadadas a proporcionar a mesma melhoria.

Contudo, no decorrer do presente texto foram apresentadas evidências de que a realidade é mais diversa do que esta abordagem retrata, embora a tradição de olhar a contribuição do RTO sob esta ótica reducionista faça com que algumas perguntas simples nunca sejam feitas pelos seus usuários. Por exemplo, quando comparamos o lucro medido no início de dois ciclos consecutivos do RTO e constatamos que o lucro aumentou, uma pergunta válida seria: “Esta variação de lucro se deu *por causa* do RTO ou *apesar* do RTO?”. Em outras palavras, esta pergunta poderia ser assim colocada: “O lucro aumentou somente devido à ação benigna do RTO, de identificar uma nova oportunidade e aproveitá-la, ou houve uma mudança no cenário de operação que naturalmente aumentaria o lucro para um valor ainda maior do que o observado, mas sua amplitude da variação natural do lucro foi restringida pelo RTO?” O fato é que esta pergunta, na falta de uma análise mais objetiva, costuma ser respondida sob um viés muito otimista, creditando integralmente ao RTO a responsabilidade do ganho quando o lucro aparente aumenta e assumindo que o RTO impediu uma queda maior, quando é observado um decréscimo do lucro.

Outras perguntas que não costumam ser feitas são: “Esta variação de lucro é real?” (ou seja, como distinguir, no que é obervado por Lm, a realidade contida em L?); “A ação que o RTO implementou era necessária?” (ou seja, a motivação da mudança de **u** pode ter sido as deficiências contidas no processamento da informação do sistema?). Todas estas perguntas ficam ocultas em parte devido ao desconhecimento das causas da variabilidade do RTO (vide Seção 3.1.1.1), e em parte porque não há quaisquer

subsídios fornecidos pelos *softwares* comerciais de RTO que apoiem esta discussão por parte do usuário. Como visto na Seção 3.1.1, e resumido nas propriedades da variabilidade de sistemas de RTO apresentadas na página 71, na ausência de condições ideais de completude, processamento e fidelidade das informações, o RTO produzirá variabilidade não nula ainda que na ausência de quaisquer mudanças no cenário de operação, como sintetizado nas Equações (140-141), na página 61.

Ainda sob o tema da variabilidade, as questões de estabilidade de sistemas de RTO, conforme abordadas na Seção 3.1.1.2, não podem ser desprezadas, apesar de constituírem um tópico inexplorado na literatura técnica da área. Ao incorporar à sua entrada os efeitos de suas próprias ações anteriores, o RTO torna-se análogo, em termos de susceptibilidade, a outros sistemas que operam em malha fechada, incorporando à sua entrada os efeitos de suas próprias ações anteriores. Tal modo de atuar é análogo ao do controle regulatório, embora no caso do RTO estas questões se desdobrem em um espaço de dimensões muito mais ampliadas, além do fato de que, no RTO, os *set-points* não são inequivocamente conhecidos. A instabilidade não só adiciona variabilidade permanente ao processo como também coloca uma restrição definitiva à capacidade do sistema de, baseado no método iterativo sequencial de atualização/otimização do RTO, levar o sistema ao ponto de ótimo desempenho financeiro. É importante notar que há uma noção difusa, entre os usuários deste tipo de sistema, de que apenas desvios "grandes" entre realidade e modelo são motivos de preocupação. Contudo, dada a natureza integrada e não linear da otimização em malha fechada executada pelo RTO, mesmo mudanças sutis podem originar alterações de comportamento imprevisíveis ao olhar desavisado, o que é exemplificado no caso 2 do estudo apresentado na Seção 3.1.1.1, no qual um modelo "perfeito" inserido em um RTO no qual uma das variáveis fixas possui valor atribuído incorreto (processamento incorreto do tipo 1) é capaz de tornar o sistema instável.

Os sistemas de RTO disponíveis no mercado reúnem alguns conceitos familiares, que parecem fazer sentido se tomados separadamente (estimação de parâmetros, otimização, detecção de estado estacionário) mas que, quando integrados sob o sistema de RTO em duas etapas, só sob condições muito específicas poderão cumprir o que prometem pois esta estratégia não está preparada [45,52,114] para perseguir as condições de otimalidade sob condições de processamento incorreto da informação (vide Seção 2.2.2).

A natureza desconexa com que estas idéias são reunidas nos sistemas comerciais de RTO em dupla camada fica mas evidente ao se observar os subsistemas de detecção de estacionariedade (SSD). Como visto na Seção 3.2.3, os métodos disponíveis na literatura para este fim e aqueles empregados pelos *softwares* comerciais não estão compromissados com o desempenho global do sistema de RTO. Este descompromisso, traduzido na falta de consequência direta entre as ações tomadas pelo SSD e a finalidade última do RTO, também impõe uma tarefa ao usuário que dificilmente será bem resolvida, que é a de definir a parametrização do SSD, além da escolha das variáveis cujo comportamento deve ser investigado (índices **ee** em **Zm**). Note-se que o usuário tem de definir, sem amparo maior que a sua experiência, diversas escolhas (vide Equação 243): o tamanho da janela de dados, as variáveis a serem testadas, o tipo de teste, sua parametrização para cada variável, além de definir regras empíricas tais como a fração mínima de sinais que devem ser considerados estacionários para que todo o sistema assim o seja considerado.

Na falta de um *padrão-ouro* de estacionariedade e de uma relação de consequência das escolhas com o desempenho do RTO, o usuário tenderá a realizar estas escolhas de modo que as decisões do SSD validem sua visão subjetiva da estacionariedade, que normalmente é baseada na análise monovariável do gráfico de cada uma das variáveis de processo consideradas. Isto fica claro na Tabela 11 (pg. 183), onde é mostrada, para um RTO real, a distância entre os veredictos dados pelo teste de hipóteses original do SSD e o resultado prático, fruto da manipulação das tolerâncias do método (Eq. 353). O modo como o usuário de um RTO real atua sobre a parametrização do SSD reflete a intenção subjetiva de que esta manipulação induza a produção de resultados que validem sua própria visão conceitual do fenômeno observado. Na verdade, mesmo a literatura acadêmica é pródiga na validação intuitiva da estacionariedade presente em resultados experimentais. Para isto, recorre-se comumente à subjetividade e à anuência tácita do leitor, como pode ser observado neste trecho, retirado do artigo de Narasimhan *et alii* [86], no qual o autor identifica as regiões de estacioneriedade em um gráfico: "*... In between these extremes and over a wide range of time periods, however, there is an* ***unambiguous*** *pattern of steady states recognizable by any* ***reasonable*** *observer…*"

Contudo, como discutido na Seção 3.2.3.2, não há como prever, *a priori*, a influência da morfologia de cada sinal, tomado individualmente, sobre a magnitude das tolerâncias assumidas de forma oculta sobre as derivadas das variáveis de estado e sobre

a consequente estimação dos parâmetros do modelo, conforme mostrado na Equação (292), página 143. Além disto, a natureza essencialmente monovariável dos testes não colabora com a análise, que esconde consequências importantes e distintas para cada sinal considerado, como visto na Figura 60 (pg. 184).

Ainda mais importante, foi mostrado ao longo da Seção 3.2 que os diversos testes de estacionariedade disponíveis na literatura e usados em *softwares* comerciais olham, cada um, para sua própria definição particular de estacionariedade, focando em especificidades da morfologia dos sinais que, por fim, tornam a conclusão dependente da escolha prévia do método e de sua parametrização. Além disto, estas definições particularizam e idealizam o cenário de variabilidade (pdf do ruído, morfologia da variação não estocástica, ausência de interação entre os sinais etc...). Na verdade, a principal questão realçada na Seção 3.2.3.2 é que, se uma visão mais pragmática for assumida, criando relações de consequências bem definidas entre a configuração do SSD e o desempenho do RTO, há duas implicações importantes: 1) o conceito de detecção de estacionariedade torna-se irrelevante, na medida em que o foco deixa de ser na origem, mas no uso; 2) é necessário recorrer-se ao uso de uma representação dinâmica do processo, ainda que o RTO opere em um cenário estacionário. Isto quer dizer que, ainda que a adaptação/optimização se refiram ao subconjunto **Z**, o teste de adequabilidade dos sinais deve ser realizado no espaço que contém **ZZ**, o conjunto completo das variáveis que descrevem o processo. Esta última conclusão é certamente impactante uma vez que a opção do uso de uma representação estacionária advém de uma tentativa de fugir das dificuldades de representação, modelagem e validação de um modelo dinâmico.

Um RTO puramente baseado em uma representação estacionária, ainda que perfeito, sempre esbarrará na impossibilidade de que uma decisão com grau mensurável de adequabilidade seja tomada a respeito de seu uso. Em verdade, a opção simplificadora de otimizar estados estacionários sempre estará vulnerável também à ausência do cômputo do custo (e da viabilidade) das trajetórias na formulação do problema de otimização. Na falta de um mapeamento dinâmico e de um estudo sobre a frequência das perturbações, a otimização estacionária não tem como garantir o contínuo acréscimo do desempenho econômico da unidade industrial, a menos que o sistema possua dinâmica tão rápida que possa ser sempre considerado em estado estacionário, o que não é comum na indústria química, embora haja exceções [115].

Um dos grandes apelos do RTO é a expectativa de que, além de continuamente otimizar o lucro da operação, seja possível estimar informações não disponibilizadas

explicitamente pela instrumentação da planta. A adaptação do modelo, sob a forma do problema de reconciliação/estimação de parâmetros torna-se um subproduto de grande interesse para o diagóstico e manutenção preventiva dos equipamentos, além de potencialmente gerar conhecimento inferencial a respeito da qualidade das correntes de matéria-prima e de produtos.

Contudo, este benefício torna-se muito difícil de ser obtido em problemas reais devido a uma série de fatores, conforme as discussões contidas nas Seções 3.3 e 4.2. O método de máxima verossimilhança, expresso como a minimização do somatório do quadrado dos desvios (Equações 316, 323), pressupõe uma longa lista de requisitos (R1-R6, pg. 161) que não só é improvável de ser atendida, como também não é submetida à validação no uso prático, sendo estes requisitos assumidos como hipóteses aceitas *a priori* pelos *softwares* comerciais disponíveis na atualidade. Como exemplo, pode-se citar a negligência em caracterizar a matriz de variância-covariância, **V** (Eq. 324), ou mesmo em ao menos estimar os termos de sua diagonal principal. Na verdade, no uso prático, os valores das variâncias das medidas são usados como parâmetros de sintonia do procedimento de estimação/reconciliação.

É comum que a tarefa de compor a diagonal de **V** seja delegado ao usuário pelo vendedor do sistema, e expresse o resultado da popular pergunta: “Em que medidas você confia mais?”. Contudo, a experiência individual do operador/engenheiro da planta não lhe dá subsídios para responder a esta pergunta com foco no resultado do RTO. Na verdade, o usuário pode, no máximo, informar qualitativamente sua experiência em termos do histórico de problemas e de erros grosseiros apresentados. Ocorre que, na prática, a estimação de parâmetros é apresentada ao usuário sob credenciais estatísticas e científicas pretensamente sólidas, embora o modo com que se dê sua operacionalização vá no sentido de olhar todas as informações ‘difíceis’ como parâmetros de sintonia. Este dilema é claramente contraditório e contraproducente. Se o enfoque for utilitário na parametrização, deve-se olhar para o procedimento como um todo com o mesmo enfoque, esquecendo as pretensões estatísticas subjacentes ao método e à análise dos resultados. Se, por outro lado, o enfoque for fundamentalista, deve-se investir tempo e recursos para a avaliação experimental da matriz **V,** para a caracterização da função distribuição probabilidade dos erros e das relações de dependência autoregressiva das variáveis. A abordagem intermediária que é feita na prática pelos sistemas comerciais apenas vai na direção de unir as deficiências de ambos os enfoques, sem conjugar suas potenciais virtudes.

O resultado destes dilemas é a produção de parâmetros estimados cujos valores e variabilidade entre ciclos consecutivos estão muito distantes da expectativa gerada pelo significado físico dos parâmetros. Na prática industrial de uso de sistemas de RTO isto é muito comum, e pode ser claramente visto na Figura 65 (pg. 192) e nas Figuras 11 e 13 da referência [50], reproduzida no apêndice desta tese. Tais resultados deixam claro que os valores produzidos pelos sistemas não são os análogos das entidades correspondentes no mundo físico, mas tão somente artefatos matemáticos oriundos não só da violação dos requisitos do método de máxima verossimilhança como principalmente pela incompletude das informações, que faz acomodar, nos parâmetros estimados, toda a variabilidade não apreendida por observação direta (medições) ou indireta (adaptação do modelo), uma vez que $\mathbf{var} \not\subset (\mathbf{ms} \cup \mathbf{upd})$. As deficiências de completude e corrupção das informações (violação da Equação 91) e de formulação das relações de atribuições de informações (violação da Equação 92) são a causa silenciosa e muitas vezes intransponível da falta de conexão entre o mundo real e as estimativas.

Muitas vezes, a falha do RTO em fornecer este valioso subproduto gera alguma perplexidade nos usuários, pois, aparentemente, o fato de o modelo usado ser ‘rigoroso’ deveria ser garantia suficiente de sucesso. Esta perplexidade em face dos resultados obtidos é resultado, por um lado, da pouca importância dada, nas implementações industriais, às discussões contidas nas seções 2.3 e 3.3. Por outro lado, o ‘rigor’ do modelo nem sempre é colocado sob uma ótica objetiva. Em termos práticos, o ‘rigor’ está frequentemente mais relacionado à complexidade do que à acurácia, uma vez que esta nunca é realmente verificada e validada em toda a região operacional. Pretensamente, um modelo baseado em princípios fundamentais reproduziria fielmente a realidade. Contudo, há algumas considerações que devem ser levadas em conta:

1) a perfeição do modelo é condição necessária, mas não suficiente, para o sucesso do procedimento de estimação/reconciliação – é apenas o primeiro dos seis requisitos do método de máxima verossimilhança;

2) por mais ‘rigor’ que se coloque na descrição, ainda se lida com modelos nos quais as distribuições espaciais são ignoradas e no qual se lida com uma representação reduzida (Seção 2.5) de um problema dinâmico – os valores residuais das derivadas das variáveis de estado serão inadvertidamente incorporados nos resultados da estimação (Seção 3.2.3);

3) os modelos são pretensamente rigorosos na descrição de fenômenos de transferência de massa e energia no interior de alguns equipamentos, mas incorporam muitas relações empíricas, de acurácia baixa e muitas vezes desconhecida, para a descrição de propriedades físicas e químicas de correntes de matéria-prima e de produto. Isto é especialmente válido na indústria de refino de petróleo, onde muitas propriedades laboratoriais não podem ser convenientemente descritas em uma formulação calcada em princípios fundamentais pois expressam o resultado de testes laboratoriais de cunho eminentemente prático (ex: intemperismo, octanagem, ponto de fulgor etc...).

4) um modelo mais detalhado exige um conjunto maior de variáveis para descrevê-lo. Uma vez que a quantidade de informação disponível é restrita pela instrumentação existente, a maior parte do acréscimo na dimensão de **Z,** trazida pelo 'rigor' do modelo, estará relacionada a variáveis contidas no conjunto **atr**, relacionadas com funções de atribuição (Equação 54, pg. 31) e dependentes de informações *a priori* confiáveis;

Este último ponto merece destaque, pois está por detrás de muitas dificuldades de desempenho de sistemas de RTO. Pode-se resumir todo o problema de adaptação do modelo à transformação do conjunto de informações disponíveis Q_a (Equação 89) no conjunto ampliado Q_a^+ (Equação 90, pg. 42). Além da fidelidade do modelo, o conjunto Q_a é extremamente dependente da informação obtida por observação direta. Na verdade, a densidade e a qualidade das medidas requeridas para fazer frente aos requisitos mínimos de informação para processamento de um sistema de RTO industrial, composto por até centenas de milhares de equações, são muito mais elevados do que aqueles requeridos pela operação rotineira, assistida pelo controle regulatório. Contudo o dimensionamento do vetor **ms** de acordo com as demandas de desempenho do RTO é raro na literatura acadêmica [116], e ignorado na prática.

Um problema recorrente do RTO é a variabilidade introduzida no processo, cujas causas estão além da simples corrupção do sinal pela presença de ruído. Como mostrado na Seção 3.1.1, o processamento incorreto da informação, em seus diversos tipos, contribui para a variabilidade devido à realimentação inerente à operação em malha fechada do RTO. Como pode ser visto de forma bem ilustrativa nas Figuras 11 e 12 (pg. 78), o procedimento de ajuste baseado na manipulação dos parâmetros atualizados para

fazer com que as respostas do modelo coincidam com a realidade não está necessariamente associado a um bom desempenho do RTO. O bom ajuste local não garante nem a otimalidade, nem variabilidade reduzida, nem a estabilidade do sistema. Esta é uma séria limitação do RTO em duas etapas [40]. Embora a literatura forneça alternativas [52], elas se baseiam em um grau de intervenção na planta intolerável à aplicação prática.

De tudo isto, fica a questão: “O dinheiro e os recursos humanos investidos na compra, implementação e manutenção de sistemas de RTO são pagos pelos benefícios introduzidos pelo sistema?” Certamente não há como responder afirmativamente a esta pergunta, ao menos não sob um ponto de vista genérico. Uma vez que o RTO em duas etapas e, particularmente, os sistemas comercializados de RTO industriais, não possuem garantias intrínsecas de otimalidade das soluções propostas nem de eficiência e acurácia na estimação dos parâmetros e nem de correta seleção das janelas de estacionariedade dos sinais, não há garantia *a priori* de que se obtenham os benefícios esperados: deve-se contar com as peculiaridades do processo em questão para que colaborem de tal modo que as vulnerabilidades inerentes ao sistema não se mostrem importantes no resultado final.

É importante perceber que, em sistemas de RTO em duas camadas, com todas suas vulnerabilidades, é mais útil pensar nos parâmetros ajustados como entidades puramente matemáticas, que devem ser colocadas a serviço do desempenho mais robusto do RTO como um todo. Embora esta postura requeira certo grau de desprendimento da habitual expectativa compartilhada por usuários e provedores de serviços de RTO, o fato é que a escolha dos parâmetros estimados feita com foco no diagnóstico da operação não implica que se alcance um bom desempenho do RTO como otimizador (e nem mesmo como estimador). Pelo contrário, a escolha da atualização de parâmetros tais como eficiências de colunas de destilação, coeficientes de troca térmica e qualidade da carga, feitas puramente por sua importância para o usuário pode tornar o sistema pior, se comparado com escolhas sem apelo intuitivo. Não só a estimação de parâmetros que estejam a serviço do diagnóstico não garante ao sistema robustez face a cenários de operação variáveis como também as próprias estimativas podem não ser úteis por conta de questões de estimabilidade, ou da interação entre as duas camadas de otimização operando em malha fechada. Tal fato pôde ser exemplificado na discussão contida na Seção 5.3, onde verificou-se que algumas das melhores estruturas de RTO, supunham a

atualização de variáveis que, na planta real, permaneciam constantes ao longo de todos os cenários de operação.

Neste contexto, até mesmo práticas costumeiras de manutenção de sistemas de RTO, tais como aquelas baseadas em procedimentos denominados *super-days*, correm o risco de serem menos úteis do que se imagina. Estes procedimentos baseiam-se em um esforço de análise laboratorial, concentrado em curto período de tempo, para prover informações não disponíveis *on-line* na prática cotidiana. A idéia é que, de posse destes dados, alguns parâmetros não estimados na operação contínua possam ser atualizados, garantindo assim melhor desempenho do RTO. Contudo, as mesmas vulnerabilidades associadas aos procedimentos contínuos do RTO se aplicam a estes procedimentos eventuais, a saber: 1) não necessariamente as entidades matemáticas cuja recalibração está associada ao melhor desempenho global futuro do RTO são aquelas associadas aos parâmetros do modelo que serão recalibrados (como eficiências de coluna etc...), o que remete ao problema da escolha do conjunto **upd**; 2) não necessariamente as variáveis cujo afastamento modelo-realidade são mais importantes para o desempenho do sistema são aquelas associadas às medidas laboratoriais feitas no *superday*, o que remete o problema da escolha do conjunto **obj**; 3) o bom ajuste local das respostas do modelo não garante o melhor desempenho global do RTO nos cenários de operação, como visto na Seção 3.1.1.1.

Todas os questionamentos surgidos a partir da análise crítica desenvolvida ao longo deste trabalho apontam para um caminho onde torna-se muito difícil não questionar seriamente o uso indiscriminado do RTO como solução garantida de condução ao desempenho econômico ótimo e de diagnóstico. Contudo, nem sempre os usuários reportam insatisfação com o uso deste tipo de ferramenta. Isto se deve ao fato de que a maioria dos problemas apresentados neste trabalho seja de natureza silenciosa. Como visto na Figura 4 (pg.62), e na Figura 69 (pg. 208), o ciclo de execução do RTO produz um elevado número de informações, com diferentes graus de observabilidade e em níveis distintos de abstração. De todo o conjunto destas informações, necessárias ao completo entendimento e diagnóstico das ações do sistema, a maior parte é inacessível ao usuário. Na verdade, não há na literatura técnica uma métrica consolidada e absoluta que ampare com segurança o uso de tais sistemas. Contudo, ainda assim, o usuário de sistemas que já estão em operação deveriam ficar atentos a indícios contidos na informação produzida pelo sistema, como algumas das análises apresentadas no Capítulo 4, que ressaltam sinais de atenção principalmente em relação à incompatibilidade entre a

frequência esperada das variação das condições de contorno e mudanças na variabilidade de curto e longo prazo das funções objetivos, valores estimados e lucro. Infelizmente, os sistemas comerciais não produzem estas análises, cabendo ao usuário a coleta, análise e síntese das informações produzidas pelo RTO.

Se não estiver amparado por bons mecanismos de suporte à decisão, a vulnerabilidade que realmente incomodará o usuário diz respeito a ocorrências que interrompam a operação contínua, o que normalmente está relacionado à não convergência dos procedimentos de otimização. Nos RTOs convencionais, tais fatos estão relacionados ao método numérico comumente usado, a *Programação Quadrática Sequencial* (SQP), cujas deficiências em problemas de RTO são conhecidas na literatura [50,117,118].

Muitos dos problemas e dos custos poderiam ser evitados mediante a correta colocação do problema a ser enfrentado na operação e otimização da planta. Grande parte do esforço de implementação de projetos de RTO concentra-se na modelagem e na construção de grandes *flowsheets* e na configuração destas informações nas interfaces dos *softwares*, mas tais projetos e as soluções comerciais que os suportam não prestam atenção à robustez do projeto implementado.

Duas questões fundamentais, relativas à fase de projeto e de operação deveriam ser respondidas como pré-requisito à implementação de quaisquer sistemas, e não apenas do caso específico do RTO: 1) “Este sistema produz soluções melhores que outras alternativas mais simples e menos dispendiosas?”, 2) “É possível medir o desempenho da ferramenta durante a operação de modo a apurar se ela cumpre o que promete?” Contudo, tais perguntas, ainda que essenciais, comumente não são respondidas. Na verdade, raramente são formuladas de forma explícita.

As razões por detrás deste fato são mais de ordem cultural que técnica. Muitas vezes, quer devido à pressão pela produção de bons resultados, quer pela percepção incompleta subjacente a uma abordagem intuitiva de um problema muito complexo, forma-se um consenso ao redor de um conceito, que passa a ter um valor por si mesmo ao invés de ser medido por sua relação funcional com a realidade na qual está inserida. Tal atitude vai de encontro à noção de refutabilidade [119], que é o pressuposto filosófico da ciência moderna. Em termos práticos, isso se traduz na implementação de soluções cujos reais benefícios são aceitos *a priori*, ainda que amparados por critérios difusos e não testados. A noção de refutabilidade fica invalidada na medida em que possíveis imperfeições não servem para questionar a solução em si, mas apenas para

justificar melhorias possíveis. Este modo de proceder é expresso na noção de *gradualidade*, que definiremos, nesta tese, como as atitudes relacionadas à adoção de um valor suposto intrinsecamente bom, e para o qual evidências em sentido contrário não o refutam mas, ao contrário, o perpetuam, no sentido de que servem para estimular a busca por um aperfeiçoamento sempre considerado possível no contexto de seu uso.

A noção de gradualidade permeia o uso de diversas soluções tecnológicas e está definitivamente associado ao uso do RTO. É muito comum que as decisões de projeto assumam a hipótese de seu caráter intrinsicamente bom como justificativa para sua implementação, uma vez que seus ganhos são estimados pelo produto da carga da unidade por uma constante positiva. Além disto, uma vez implementado, possíveis inconsistências nos resultados produzidos pelo sistema não são usados para questionar a ferramenta em si, mas sim para justificar reajustes de parâmetros de sintonia, realizações de *superdays* etc..

Contudo, como visto no exemplo simples mostrado na Figura 5, pg. 70, o RTO não é *inerentemente bom*, uma vez que, como mostrado neste caso, o sistema estava em seu ponto ótimo e foi dele removido a partir do instante em que o RTO foi ligado. Além disto, pode haver soluções muito mais simples que resultem em desempenho superior. Tomando-se como exemplo o estudo de caso apresentado no Capítulo 5, um exercício interessante seria descobrir se existem escolhas das variáveis de decisão que, mantidas fixas, fossem capazes de levar o sistema a desempenhos melhores do que aqueles obtidos sob a ação de um RTO. A resposta a esta pergunta, calculada sobre a região formada pelas 15 condições operacionais previamente definidas, pode ser vista na Figura 83. Nesta Figura, a área assinalada na cor branca indica pares das variáveis de decisão F_B e T que, ainda que mantidos constantes, produzem melhor desempenho relativo da métrica $\mathcal{M}$luv (Eq. 421) do que o RTO configurado com a melhor estrutura disponível. Note-se que $\mathcal{M}$luv é a versão não normalizada de mluv (Seção 5.3).

$$\mathcal{M}luv = -\mathcal{M}L + \mathcal{M}u + \mathcal{M}V \qquad (421)$$

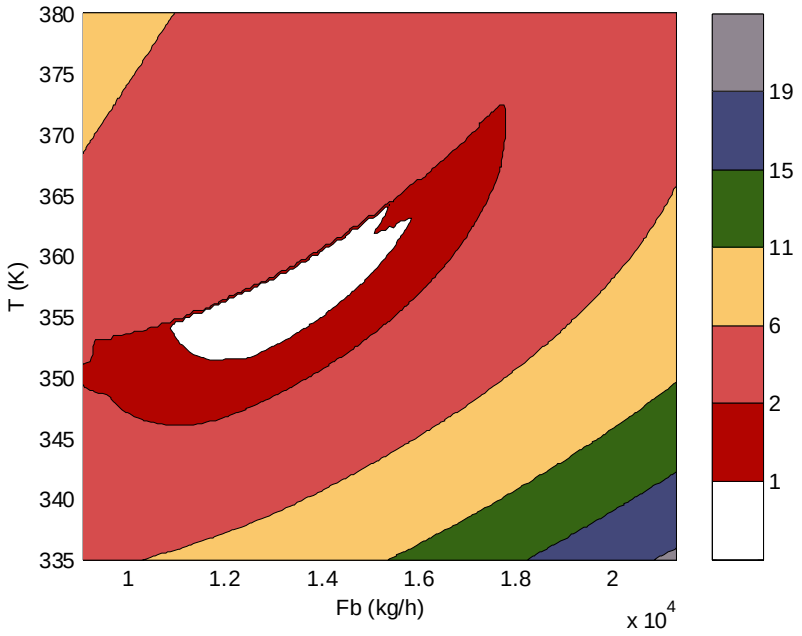


Figura 83 – Razão entre a métrica $\mathcal{M}luv$ para um sistema sem RTO e para um sistema com RTO configurado com a sua melhor estrutura. Processo descrito no estudo de caso do Capítulo 5, na versão *restr* do problema. A região assinalada com a cor branca (razão menor que 1) indica desempenho melhor do sistema sem RTO.

A exigência de um modelo dinâmico para a detecção da adequabilidade é um obstáculo irreconciliável com a formulação padrão do RTO em duas etapas e para o qual não há alternativa viável: sistemas que se baseiem na identificação de estacionariedade como requisito para que um ciclo de otimização seja implementado e que não disponham de uma representação dinâmica do processo estão fadados a decisões desconectadas de suas consequências reais. Contudo, ainda que isto seja esquecido, alguns requisitos não atendidos pelos projetos atuais de RTO deveriam sê-lo de modo a proporcionar alguma garantia de que o sistema cumpra o que dele se espera:

- Uso de métricas de desempenho operacional que quantifiquem de forma *realista* os benefícios obtidos baseados nas informações *disponíveis* para o usuário durante a operação cotidiana;

- Avaliação dos cenários de operação (variações das perturbações ao longo do tempo, em magnitude e frequência; análise objetiva da frequência de intervenção do RTO; inclusão de vulnerabilidades da instrumentação);

- Estudo do comportamento em malha fechada sob a presença de incertezas

- Escolha da estrutura do RTO sob um ponto de vista integrado, exaustivo e não intuitivo

- Previsão da modificação da estrutura escolhida caso quaisquer premissas de operação sejam modificadas

- Dimensionamento da densidade e qualidade da instrumentação requerida como parte do projeto do RTO (inclusão da definição de **ms** no problema de definição da estrutura do RTO)

Em resumo, o panorama traçado neste texto para as aplicações de otimizadores em tempo real em duas etapas infelizemente não é auspicioso. O conjunto de ações e informações necessárias para a garantia de um desempenho previsível é muito mais elevado do que aquele contido nas premissas dos projetos e implementações. Embora sempre possa haver condições específicas que o justifiquem (como o caso *semrestr*, cap. 5), tais condições só podem ser identificadas a partir de estudo exploratório exaustivo, e não em decorrência de uma decisão baseada na natureza *benigna* intrínseca ao RTO. Contudo, pelo lado positivo, é interessante ressaltar que outras soluções com menos pretensões à sofisticação não deveriam ser relegadas a segundo plano, tais como a escolha de condições robustas de operação ou mesmo a otimização *off-line*, amparada por bons recursos de análise. Tais estratégias mais simples tornam-se mais competitivas na medida em que os custos das vulnerabilidades do RTO sejam incorporados na tomada de decisão.

7. Conclusões

Esta tese apresentou o problema de RTO em duas etapas sob o arcabouço de seus fundamentos teóricos, confrontando-a com a operação cotidiana na indústria. Foram ressaltadas diversas vulnerabilidades associadas ao uso de ferramentas comerciais disponíveis para a indústria e constatado que não há garantias de que a promessa de que as oportunidades de ganho econômico sejam aproveitadas. Também foram apresentadas análises de um sistema industrial de larga escala atualmente em operação.

Os sistemas de RTO em duas etapas podem ser úteis, mas apenas sob condições específicas do processo que mitiguem suas vulnerabilidades intrínsecas. Para isto, deve ser investido esforço objetivo na análise exaustiva das configurações da estrutura do sistema. Como exposto no trabalho, apesar de as escolhas associadas à estrutura do RTO serem baseadas em decisões intuitivas, este método não está vinculado às melhores decisões.

O RTO deve ser encarado sob o enfoque de sua finalidade última: a obtenção dos melhores desempenhos financeiros possíveis. Nesta busca, em virtude de suas vulnerabilidades, devem ser abandonadas as pretensões ao uso dos parâmetros estimados com fins ao diagnóstico da operação.

Se o sistema de RTO for baseado em representações estacionárias do processo, não há como prescindir de um modelo dinâmico que ampare a decisão relacionada à autorização de seu funcionamento. Deve-se abandonar a idéia de que esta decisão seja feita com base na *detecção* da estacionariedade e incorporar a idéia de teste de *adequabilidade* multivariável, focada nas consequências da decisão.

As idéias aqui expostas reforçam o requisito de que qualquer sistema candidato ao uso industrial só deva ser implementado se houver uma métrica *realista* que possa comprovar seu desempenho. Para o RTO em duas etapas, o cálculo diligente dos custos associados a suas vulnerabilidades, se incluído nos critérios associados à seleção de tecnologias de otimização, pode indicar que tecnologias com menor pretensão à sofisticação, como a atualização *off-line* ou o aproveitamento de condições auto-otimizáveis se mostrem mais competitivas quando observado o balanço entre os benefícios produzidos e dispêndio de recursos financeiros e humanos em sua implementação e operação.

8. Referências Bibliográficas

[1] WHITE, D. C., “Online Optimization: What Have We Learned?”, Hydrocarbon Processing, **7**7, pp. 55–59, 1998.

[2] DARBY, M., M. NIKOLAOU, J. JONES AND D. NICHOLSON, “RTO—An Overview and Assessment of Current Practice”, J. Process Contr. 21, 874–884 (2011).

[3] FRIEDMAN, Y. Z., ‘‘Closed-Loop Optimization Update—A Step Closer to Fulfilling the Dream,’’ Hydrocarbon Processing, **7**9, pp. 15–16 (2000).

[4] ZHANG, Y.; FORBES, F., "Extended design cost - a performance criterion for real-time optimization systems", *Computers and Chemical Engineering*, v. 24, pp. 1829-1841, 2000.

[5] BOX, G.E.P., “On The Experimental Attainment of Optimum Conditions”, *Journal of the Royal Statistical Society B*, pp. 1-45, 1951.

[6] BOX, G.E.P., “The Exploration and Exploitation of ResponseSurfaces: Some General Considerations and Examples”, *Biometrics*, pp. 16-60, 1954

[7] BAMBERGER, W.; ISERMAN, R., “Adaptive On-Line Steady State Optimization of Slow Dynamic Processes”, *Automatica*, v. 14, pp. 223-230, 1978

[8] GARCIA, C.E.; MORARI, M., “Optimal Operation of Integrated Processing Systems”, *AICHE Journal*, v. 27, p. 960, 1981.

[9] MCFARLANE, R.C.; BACON, D.W., “Empirical Strategies for Open-Loop On-Line Optimization”, *Canadian Journal of Chemical Engineering*, v.67, pp.665-677, 1989.

[10] YIP, W.S.; MARLIN, T.E., "The effect of model fidelity on real-time optimization performance", *Computers and Chemical Engineering*, v. 28, pp. 267-280, 2004.

[11] SKOGESTAD, S., “Control structure design for complete chemical plants”, *Computers and Chemical Engineering*, v. 28, pp. 219-234, 2004.

[12] ZHANG, Y.; FORBES, J.F., “Perfomance Analysis of Perturbation-Based Methods for Real-Time Optimization”, *Canadian Journal of Chemical Engineering*, v.84, pp. 209-218, 2006.

[13] CHACHUAT, B.; SRINIVASAN, B.; BONVIN, D., "Model Parameterization Tailored to Real Time Optimization", *Computer Aided Chemical Engineering*, v. 25, pp. 1-13, 2008.

[14] CHACHUAT, B.; SRINIVASAN, B.; BONVIN, D., "Adaptation strategies for real-time optimization", *Computers and Chemical Engineering*, v. 33, pp. 1557-1567, 2009

[15] CHEN, C.Y.; JOSEPH, B., "On-Line Optimization Using a Two-Phase Approach - An Application Study", *Ind. Eng. Chem. Res.*, v. 26, pp. 1924-1930, 1987

[16] ROBERTS, P. D., “An algorithm for steady-state system optimization and parameter estimation”, *International Journal of Systems Science*, v.10, pp. 719-734, 1979.

[17] FORBES, J.F.; MARLIN, T.E., "Model Accuracy for Economic Optimizing Controllers - The Bias Update Case", *Industrial Engineering Chemical Research*, v. 33, pp. 1919-1929, 1994

[18] TATJEWSKI, P., “Iterative optimizing set-point control—The basic principle redesigned”, Proceedings of the 15th Triennial IFAC World Congress, 2002

[19] GAO, W.; ENGELL, S., "Iterative set-point optimization of batch chromatography", *Computers and Chemical Engineering*, v. 29, pp. 1401-1409, 2005

[20] SKOGESTAD, S.; “Self-optimizing control: The missing link between steadystate optimization and control”, *Computers and Chemical Engineering*, v. 24, pp. 569–575, 2000.

[21] FRANÇOIS, G., SRINIVASAN, B., & BONVIN, D., “Use of measurements for enforcing the necessary conditions of optimality in the presence of constraints and uncertainty”, *Journal of Process Control*, v. 15, pp. 701–712, 2005.

[22] SRINIVASAN, B., BIEGLER, L. T., & BONVIN, D., “Tracking the necessary conditions of optimality with changing set of active constraints using a barrier-penalty function”, Computers and Chemical Engineering, v. 32, pp. 572–579, 2008.

[23] NAYSMITH, M.R.; DOUGLAS, P. L., “Review of Real Time Optimization in the Chemical Process Industries”, *Asia-Pacific Journal of Chemical Engineering*, v. 3, pp.67-87, 2008.

[24] MILETIC, I.; MARLIN, T., "Results analysis for real-time optimization (RTO): Deciding when to change the plant operation", *Computers and Chemical Engineering*, v. 20, pp. 1077-1082, 1996.

[25] MILETIC, I.P.; MARLIN, T.E., “On-line Statistical Results Analysis in Real-Time Operations Optimization”, *Ind. Eng. Chem. Res.*, v. 37, pp. 3670-3684, 1998.

[26] MILETIC, I.P.; MARLIN, T.E., "Results diagnosis for real-time process operations optimization", *Computers and Chemical Engineering*, v. 22, pp. 8475-8482, 1998.

[27] ZHANG, Y.; NADLER, D.; FORBES, J.F., "Results analysis for trust constrained real-time optimization", *Journal of Process Control*, v. 11, pp. 329-341, 2001.

[28] LOEBLEIN, C.; PERKINS, J.D.; SRINIVASAN, B.; BONVIN, D., "Economic performance analysis in the design of on-line batch optimization systems", *Journal of Process Control*, v. 9, pp. 61-78, 1999

[29] ZHANG, Y.; MONDER, D.; FORBES, J.F., "Real-time optimization under parametric uncertainty: a probability constrained approach", *Journal of Process Control*, v. 12, pp. 373-389, 2002.

[30] DARLINGTON, J.; PANTELIDES, C. C.; RUSTEM, B.; TANYI, B. A., "An algorithm for constrained nonlinear optimization under uncertainty", *Automatica*, v. *35*, pp.217-228, 1999.

[31] KALL, P.; WALLACE, S.W., *Stochastic Programming*, John Wiley & Sons, Chichester, 1994.

[32] OSTROVSKY, G.M.; ZIYATDINOV, N.N.; LAPTEVA, T.V., "One-stage optimization problem with chance constraints", *Chemical Engineering Science*, v. 65, pp. 2373-2381, 2010.

[33] KOOKOS, I.K., "Optimal Operation of Batch Processes under Uncertainty: A Monte Carlo Simulation-Deterministic Optimization Approach", *Ind. Eng. Chem. Res.*, v. 42, pp. 6815-6822, 2003.

[34] MULVEY, J.M.; VANDERBEI, R.J., "Robust optimization of large-scale systems", *Operations Research*, v. 43, pp. 264–281, 1995.

[35] LI, P.; ARELLANO-GARCIA, H.; WOZNY, G., "Chance constrained programming approach to process optimization under uncertainty", *Computers and Chemical Engineering*, v. 32, pp. 25-45, 2008.

[36] MESFIN, G.; SHUHAIMI, M., "A chance constrained approach for a gas processing plant with uncertain feed conditions", *Computers Chemical Engineering*, v. 34, pp. 1256-1267, 2010.

[37] LOEBLEIN, C.; PERKINS, J.D., "Economic Analysis of Different Structures of on-line Process Optimization Systems", *Computers and Chemical Engineering*, v.20, pp. 551-556, 1996.

[38] KRISHNAN, S.; BARTON, G.W.; PERKINS, J.D., "Robust Parameter Estimation in on-line optimization - part I. Methodology and Simulated Case Study", *Computers and Chemical Engineering*, v. 16, pp. 545-562, 1992.

[39] KRISHNAN, S.; BARTON, G.W.; PERKINS, J.D., "Robust Parameter Estimation in on-line optimization - part 2. Application to an Industrial Process", *Computers and Chemical Engineering*, v. 17, pp. 663-669, 1993.

[40] FORBES, J.F.; MARLIN, T.E.; MACGREGOR, J.F., "Model adequacy requirements for optimizing plant operations", *Computers and Chemical Engineering*, v. 18, pp. 497-510, 1994.

[41] GANESH, N.; BIEGLER, L. T., "A reduced Hessian strategy of sensitivity analysis of optimal processes", *AIChE Journal, v. 33, pp. 282-296, 1987.*

[42] LOEBLEIN, C.; PERKINS, J., "Economic analysis of different structures of on-line process optimization systems", *Computers Chemical Engineering*, v. 22, pp. 1257-1269, 1998.

[43] LOEBLEIN, C.; PERKINS, J., "Structural Design for On-Line Process Optimization - I. Dynamic Economics of MPC", *AICHE Journal*, v. 45, pp. 1018-1029, 1999.

[44] de HENNIN, S.R.; PERKINS, J.D.; BARTON, G.W., "Structural decisions in on-line optimization", *Proc. International Conference on Process Systems EngineeringPSE '94*, 297-302, 1994.

[45] FORBES, J.F.; MARLIN, T.E., "Design cost: a systematic approach to technology selection for model-based real-time optimization systems", *Computers and Chemical Engineering*, v. 20, pp. 717-734, 1996.

[46] PINTO, J.C., "On the costs of parameter uncertainties. Effects of parameter uncertainties during optimization and design of experiments", *Chemical Engineering Science*, v. 53, pp. 2029-2040, 1998.

[47] GATTU, G.; PALAVAJJHALA, S.; ROBERTSON, D.B., "Are oil refineries ready for non-linear control and optimization?", International Symposium on Process Systems Engineering and Control, Mumbai, 2003.

[48] YIP, W.S.; MARLIN, T.E., "Multiple data sets for model updating in real-time operations optimization", *Computers and Chemical Engineering*, v. 26, pp. 1345-1362, 2002.

[49] YIP, W.S.; MARLIN, T.E., "Designing plant experiments for real-time optimization systems", *Control Engineering Practice*, v. 11, pp. 837-845, 2003.

[50] QUELHAS, A.D.; JESUS, N.J.C.; PINTO, J.C., "Common Vulnerabilities of RTO Implementations in Real Chemical Processes", *Canadian Journal of Chemical Engineering*, v.91, pp. 652-668, 2013.

[51] SCHWAAB, M.; PINTO, J.C.C., *Análise de Dados Experimentais I*, E-papers, 2007.

[52] MARCHETTI, A.; CHACHUAT, B.; BONVIN, D., "Modifier-Adaptation Methodology for Real-Time Optimization", *Industrial Engineering Chemical Research*, v. 48, pp. 6022-6033, 2009.

[53] ROLANDI, P.A.; ROMAGNOLI, J.A.; "*Simultaneous Dynamic Validation/Identification of Mechanistic Process Models and Reconciliation of Industrial Process Data*", Computer Aided Chemical Engineering, v.21, pp. 267-272, 2006.

[54] BRITT, H. I.; LUECKE, R. H. "The estimation of parameters in nonlinear, implicit models", *Technometrics*, v. 15, pp. 233-247, 1973.

[55] KIM, I.; LIEBMAN, M. J.; EDGAR, T. F. "Robust error-in-variables estimation using nonlinear programming techniques", *American Institute of Chemical Engineering Journal*, v. 36, pp. 985-993, 1990.

[56] KIM, I.; EDGAR, T. F.; BELL, N. H. "Parameter estimation for a laboratory water–gas-shift reactor using a nonlinear error-invariables method", *Computers and Chemical Engineering*, v. 15, pp. 361-367, 1991.

[57] MACDONALD, R. J.; HOWAT, C. S. "Data reconciliation and parameter estimation in plant performance analysis", *American Institute of Chemical Engineering Journal*, v. 34, pp. 1-8, 1988.

[58] BIEGLER, L.T.; GROSSMAN, I.E.; WESTERBERG, A.W., "A note on approximation techniques used for process optimization", *Computer and Chemical Engineering Applications of Artificial Intelligence*, v. 9, pp. 201-206, 1985.

[59] GRIFFEL, D.H., *Applied Functional Analysis*, Dover Publications, 2002.

[60] KREYSZIG, E., *Advanced Engineering Mathematics*, John Wiley and Sons, 2006.

[61] KREYSZIG, E., *Introductory Functional Analysis*, John Wiley and Sons, 1989.

[62] ROBERTS, P.D.; WILLIAMS, T.W.C., "On an algorithm for combined system optimisation and parameter estimation", *Automatica*, v. 17, pp. 199-209, 1981.

[63] AVRIEL M., Nonlinear *Programming: Analysis and Methods.* Prentice-Hall, New Jersey, 1976.

[64] BRDYŚ, M.; ELLIS, J.E., ROBERTS, P.D., "*Augmented integrated system optimization and parameter estimation technique: Derivation, optimality and convergence"*, IEEProc.-D, v. 134, pp.201-209, 1987.

[65] CHACHUAT, B.; MARCHETTI, A.; BONVIN, D. Process optimization via constraints adaptation. *J. Process Control*, v. 18, pp. 244–257, 2008.

[66] FLETCHER R., *Practical Methods of Optimization*, Wiley, New York, 1987.

[67] KEDEM, B.; FOKIANOS, K.,*Regression Models for Time Series Analysis.*1 ed. Wiley-Interscience, New York, 2002.

[68] JIANG, T.; CHEN, B.; HE, X.; STUART, P., "Application of steady-state detection method based on wavelet transform", *Computers and Chemical Engineering*, v. 27, pp. 569-578, 2003.

[69] FATH, B.D.; CABEZAS, H.; PAWLOWSKI, C.W., "Regime changes in ecological systems: an information theory approach", *Journal of Theoretical Biology*, v.222, pp. 517-530, 2003.

[70] ALEKMAN, S.L., *Control for the Process Industries*, Putman Publications, Chicago, IL, v. 7, número 11, pp. 62, novembro 1994.

[71] SCHLADT, M.; HU, B., "Soft sensors based on nonlinear steady-state data reconciliation in the process industry", *Chemical Engineering and Processing*, v. 46, pp. 1107-1115, 2007.

[72] JUBIEN, G.; BIHARY, G.*Control for the Process Industries*, Putman Publications, Chicago, IL., v. 7,número 11, pp.64, novembro 1994

[73] KIM, M.; YOON, S.H.; DOMANSKI, P.A.; PAYNE, W.V., "Design of a steady-state detector for fault detection and diagnosis of a residential air conditioner", *International Journal of Refrigeration*, v. 31, pp. 790-799, 2008.

[74] MAHULI, S.K.; RHINEHART, R.; RIGGS, J.B., "Experimental demonstration of non-linear model-based in-line control of pH", *Journal of Process Control*, v. 2, pp. 145-153, 1992.

[75] MORENO, R.P., *Steady State Detection, Data Reconciliation, and Gross Error Detection - Development for Industrial Processes*, M.Sc., University of New Brunswick, 2010.

[76] ÖNÖZ, B.; BAYAZIT, M., "The Power of Statistical Tests for Trend Detection", *Turkish Journal of Engineering & Environmental Sciences*, v.27, pp.247-251, 2003.

[77] MONTGOMERY, D.C.; RUNGER, G.C., *Applied statistics and probability for engineers.* 3ª ed. John Wiley & Sons, 2002.

[78] VON NEUMANN, J.; KENT, R.; BELLINSON, H.; HART, B., "The mean square successive difference", *Ann. Math. Stat.*, pp. 153-162, 1941.

[79] YOUNG, L.C., "On Randomness in Ordered Sequences", *The Annals of Mathematical Statistics*, v.12, pp. 293-300, 1941.

[80] VON NEUMANN, J., "Distribution of the ratio of the mean square successive difference to the variance", *Ann. Math. Stat.*, pp. 367-395, 1941.

[81] CAO, S.; RHINEHART, R.R., "An efficient method for on-line identification of steady state", *Journal of Process Control*, v. 5, pp. 363-374, 1995.

[82] SHROWTIA, N.A.; VILANKARA, K.P., RHINEHART, R.R., "Type-II critical values for a steady-state identifier", *Journal of Process Control*, v. 20, pp. 885-890, 2010.

[83] BHAT, S.A.; SARAF, D.N., "Steady-State Identification, Gross Error Detection, and Data Reconciliation for Industrial Process Units", *Ind. Eng. Chem. Res.*, v. 43, pp. 4323-4336, 2004.

[84] BROWN, P.R.; RHINEHART, R.R., "Development and demonstration of a method for automated steady-state identification in multivariable systems", *Hydrocarbon Processing*, v. 79, 79–83, 2000.

[85] MANSOUR, M.; ELLIS, J.E., "Methodology of on-line optimisation applied to a chemical reactor", *Applied Mathematical Modelling*, v. 32, pp. 170-184, 2008.

[86] NARASIMHAN, S.; MAH, R.S.H.; TAMHANE, A.C.; WOODWARD, J.W.; HALE, J.C., "A Composite Statistical Test for Detecting Changes of Steady States", *AIChE Journal*, v. 32, pp. 1409-1418, 1986.

[87] NARASIMHAN, S.; KAO, C.S.; MAH, R.S.H., "Detecting Changes of Steady States Using the Mathematical Theory of Evidence", *AIChE Journal*, v. 33, pp. 1930-1932, 1987.

[88] SHAFER G. *A Mathematical Theory of Evidence*, Princeton University Press. Princeton. N J, 1976.

[89] GOURÉVITCH, B.; EGGERMONT, J.J., "A simple indicator of nonstationarity of firing rate in spike trains", *Journal of Neuroscience Methods*, v. 163, pp. 181-187, 2007.

[90] LEHMANN, E. *Nonparametrics Statistical Methods Based on Ranks*, Springer. 2006.

[91] HIPEL, K.W.; McLEOD, A.D.; *Time Series Modelling of Water Resources and Environmental Systems, Developments in Water Science*, v. 45, pp. 853-938. Elsevier. 1994.

[92] FAHIDY, T., "Potential applications of rapid/elementary nonparametric statistical techniques (NST) to electrochemical problems", *Electrochimica Acta*, v. 54, pp. 6949-6953, 2009.

[93] FLEHMIG, F.; WATZDORF, R.; MARQUARDT, W., “Identification of trends in process measurements using the wavelet transform”, *Computers and Chemical Engineering, v. 22*, pp. 491-496, 1998.

[94] FLEHMIG, F.; MARQUARDT, W., "Detection of multivariable trends in measured process quantities", *Journal of Process Control*, v. 16, pp. 947-957, 2006.

[95] CASTRO, A.; ALMEIDA, F.G; AMORIM, P.; NUNES, C.S., "A Wavelet Based Method for Steady-State Detection in Anesthesia", *Proceedings on 31st Annual International Conference of the IEEE EMBS*, , pp. 954-957, 2009.

[96] CAUMO, L.; TRIERWEILER, J.O., “Steady-state detection for multivariate systems based on PCA and wavelets”, Seminário do Programa de Pós-Graduação em Engenharia Química da Universidade Federal do Rio Grande do Sul, 2005.

[97] MASSEY, F.J., “The Kolmogorov-Smirnov Test for Goodness of Fit”, *Journal of the American Statistical Association*, v. 46, pp. 68-78, 1951.

[98] POULIN, E.; HODOUIN, D.; LACHANCE, L., "Impact of plant dynamics on the performance of steady-state data reconciliation", *Computers and Chemical Engineering*, v. 34, pp. 354-360, 2010.

[99] FLEHMIG, F.; MARQUARDT, W., "Inference of multi-variable trends in unmeasured process quantities", *Journal of Process Control*, v. 18, pp. 491-503, 2008.

[100] MALLAT, S. G.; ZHANG, Z., “Matching Pursuits with Time-Frequency Dictionaries”, *IEEE Transactions on Signal Processing*, pp. 3397–3415, 1993.

[101] MALLAT, S., *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, 3rd ed., 2009.

[102] GABOR, D., “Theory of Communication”, *J. IEE*, v. 93 (III), pp. 429-457, 1946.

[103] BAKSHI, B.R.; STEPHANOPOULOS, G., “Representation of Process Trends – III. Multiscale Extraction of Trends from Process Data”, *Computers and Chemical Engineering*, pp. 267-302, 1994.

[104] ADDISON, P.S., *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*, Taylor & Francis, 2002.

[105] MALLAT S., A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Pattern Anal. and Machine Intelligence*, v. 11, ,pp.674-693, 1989

[106] BATES, D.M.; WATTS, D.G., *Nonlinear Regression Analysis and its Applications*, John Wiley and Sons, 1988.

[107] SEBER, G.A.F.; WILD, C.J., *Nonlinear Regression*, John Wiley and Sons, 1989.

[108] HAMILTON, J.D., *Time Series Analysis*, Princeton University Press, 1994

[109] BERKSON, J., "Are There Two Regressions?", *Journal of the American Statistical Association*, pp. 164-180, 1950.

[110] FROST, C.; THOMPSON, S.G., "Correcting for regression dilution bias: comparison of methods for a single predictor variable", *J.R. Statist. Soc. A.*, pp.173-189, 2000.

[111] DRAPER, N.R.; SMITH, H.; *Applied Regression Analysis*, Wiley-Interscience, 1998.

[112] WEEKS, M.; *Digital Signal Processing Using Matlab and Wavelets*, Infinity Science Press LLC, 2007.

[113] WILLIAMS, T.J.; OTTO, R.E., "A Generalized Chemical Processing Model for the Investigation of Computer Control", *A.I.E.E Trans.*, v. 79, pp. 458-473, 1960.

[114] FRANÇOIS; G. BONVIN, D., "Use of Convex Model Approximations for Real-Time Optimization via Modifier Adaptation", *Industrial & Engineering Chemistry Research*, 52, v. 33, 11614-11625, 2013.

[115] JESUS, N. J. C., 2011, *Otimizacão em Tempo Real em um Processo Industrial de Produção de Etileno*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, Rio de Janeiro, Brasil.

[116] FRALEIGH, L., M. GUAY AND J. F. FORBES, ''Sensor Selection for Model-Based Real-Time Optimization: Relating Design of Experiments and Design Cost,'' *J. Process Contr.* 13, pp. 667–678, 2003.

[117] CUBILLOS, F. A.; ACUÑA, G.; LIMA, E. L., ''Real-Time Process Optimization Based on Grey-Box Neural Models'', *Braz. J. Chem. Eng.*, v. 3, pp. 433–443, 2007.

[118] GOLSHAN, M., M. R. PISHVAIE AND R. B. BOOZARJOMEHRY, ''Stochastic and Global Real Time Optimization of Tennessee Eastman Challenge Problem,'' Eng. Appl. Artif. Intel. 21, 215–228, 2008.

[119] POPPER, K., *A Lógica da Pesquisa Científica*, Cultrix, 1993.

9. Apêndice

Artigo “Common Vulnerabilities of RTO Implementations in Real Chemical Processes”, QUELHAS, A.D.; JESUS, N.J.C.; PINTO, J.C., *Canadian Journal of Chemical Engineering*, v.91, pp. 652-668, 2013.

COMMON VULNERABILITIES OF RTO IMPLEMENTATIONS IN REAL CHEMICAL PROCESSES

André D. Quelhas†, Normando José Castro de Jesus†† and José Carlos Pinto*

Programa de Engenharia Química/COPPE, Universidade Federal do Rio de Janeiro, Cidade Universitária. CP: 68502, Rio de Janeiro, 21945-970 RJ, Brazil

Real-time optimisation (RTO) systems face challenging scenarios in industrial practice, such as incomplete and corrupted process information, uncertain large-scale mathematical models and numerical optimisation issues. Proper design of RTO structure and robust diagnosis tools are keys for good performance, although they are neglected in commercial RTO software and not fully solved in the technical literature. This article reviews the concepts behind the two-step RTO approach and suggests performance metrics. It also points out the large list of structural decisions and the consequences of intuitive, experience-based RTO design choices. The discussed vulnerabilities are illustrated with some simulations and real industrial implementations.

Keywords: real-time optimisation (RTO), modelling, numerical methods, parameter estimation

INTRODUCTION

Industrial chemical processes are complex arrangements of a myriad of equipment and pipelines, where streams of variable compositions are transformed into useful products. Due to this inherent complexity, the act of selling a product is the materialisation of a huge collection of decisions. The key for success is the careful coordination of all possible alternatives, focussing on maximum performance in terms of profit, safety and reliability. However, the asynchronous nature of the decision-making process often makes coordination very difficult. A very large set of *irreversible* decisions, most of them hardware-related, are made even before the plant start-up, including the geometry of equipment and pipes, selection of materials and so forth. These preliminary decisions shape the degree of freedom for the posterior routine plant operation.

In common industrial practice, automated systems (mainly the regulatory control protocols) are in charge of most short time routine decisions related to the rejection of disturbances and set point tracking. If some degrees of freedom are left unused by the lower automation levels, it is possible to manage them in order to actively pursue the best profit performance along the operation time.

On the other hand, it is far less common to find real-time optimisation (RTO) systems in industry. This is justified, in part, by the fact that the implementation of RTO procedures may not be suitable for every process (White, 1998). Besides, although the idea behind RTO is very easy to understand and accept (the process operation must be optimised in real time, as the boundary conditions and relevant process parameters change), RTO systems are not fully accepted in industry (Darby et al., 2011). This is due to the fact that many real world implementations are shown to be "labor-intensive, difficult to develop and f[a]ll apart easily" (Friedman, 2000).

As a matter of fact, RTO constitutes a broad concept. In this text, it is defined as the automated implementation of business-focussed decisions, based on rigorous non-linear process models and repeated more frequently on average than the occurrence of disturbances that drive the process to suboptimal performance. In chemical processes, RTO is responsible for translating a product recipe from the scheduling layer into the *best* set of reference values to the model predictive control (MPC) layer.

†André D. Quelhas's present address is Petrobras—Petróleo Brasileiro SA, Brazil.

††Normando José Castro de Jesus's present address is Braskem SA, Brazil.

**Author to whom correspondence may be addressed.*
E-mail address: pinto@peq.coppe.ufrj.br
Can. J. Chem. Eng. 91:652–668, 2013

DOI 10.1002/cjce.21738
Published online 24 September 2012 in Wiley Online Library (wileyonlinelibrary.com).

By far, the commonest RTO scheme is the two-step approach (Chen and Joseph, 1987) used in typical commercial software (ROMeo 5.1-Invensys, Houston, TX, Aspenplus 7.1-Aspentech, Burlington, MA). For this reason, most of the RTO technical literature is somehow related to this theme. The technique owes its popularity to the intuitive idea behind it. In the first optimisation layer, plant information is used to update model parameters based on the best fitting of measurements. Next, the optimisation layer produces a set of decision variable values that are assumed to lead process to its best economic performance.

It is important to emphasise that manipulation of model parameters in order to fit available process measurements does not necessarily guarantee the construction of an adequate model for process optimisation (Forbes et al., 1994). For this reason, some alternative procedures have been proposed (Chachuat et al., 2009) based on stronger mathematical requirements and constraints that guarantee optimality of process operation. These procedures demand a series of time-consuming experimental measurements in order to evaluate gradients of a large set of functions and variables. Given the considerable impact on productivity, these implementations are virtually absent in current industrial practice. In fact, commercial software is usually based on very standard two-step RTO structure and does not even take into account collateral improvements of this approach, such as input excitation design (Yip and Marlin, 2003) or automated diagnosis (Zhang et al., 2001).

The implementation of RTO projects demands several months of specialised work. Such projects normally produce a complex mathematical system that encompasses up to hundreds of thousands of equations. Under the pressure of daily routine, it is very common that users get immersed in software implementation issues rather than in diagnosing and criticising the obtained results and software tools. Although it is possible to find some valuable criticism about RTO implementations in the open literature (Friedman, 2000; Gattu et al., 2003), this discussion is normally presented in general terms, making it hard for practitioners to distinguish process-related features from methodological limitations of the RTO approach.

The present article is intended to highlight some of the vulnerabilities related to the RTO structure commonly found in commercial software by revisiting the ideas over which the two-step RTO approach is built. Although some of the vulnerabilities described here are supposed to be known by those who work in the field, an integrated discussion about the many aspects that may be responsible for misleading interpretation and performance degradation of RTO schemes in real industrial implementations is still missing in the literature.

PROBLEM STATEMENT

The main task of an optimisation system is to tune the vector of available degrees of freedom of a process in order to reach the "best" value of some performance metric. This vector is a subset of a larger set of input variables, I, that dictates how the process behaves, as described by the set of output variables, O. In a set of algebraic equations, distinction between input and output variables is not relevant, although the real process cannot be changed instantaneously from one state to another when vector I is modified. However, as the dynamic behaviour of process variables is ruled by sets of differential equations, output variables show higher dependence on previous states than input variables, although such distinction disappears if one is focussed only on stationarity.

Input and output variables are related through a set f of equations that express conservation balances (mass, energy, momentum), equipment design constraints and so forth. Another set g of inequality relationships describes allowable operational regions according to safety conditions, product specifications and equipment operation requirements. f and g represent the so-called process model.

The input set I can commonly be partitioned into a set of stimuli and a set of parameters. Partitioning is not a mathematical requirement but can be useful to incorporate some previous physical knowledge into the model representation of the real process. As a matter of fact, it may be hard to unambiguously define both sets in a way that can be generalised to any thinkable process. Distinction between these two sets is usually based on measurability, time scale of change and physical interpretation of process, among other criteria. In this text, the set of input variables, I, will not be partitioned, and all elements of I will be treated similarly. The word "parameter" will be used to designate an offset modifier in the context of model updating.

In an ideal scenario, the RTO implementation relies on perfect knowledge of the input set I and of the mapping of $\mathrm{I} \rightarrow \mathrm{O}$ (process model) and $(\mathrm{I,O}) \rightarrow \mathrm{L}$ (performance index). As stated above, optimisation is the act of selecting the set of variables u that conducts the process to the most favourable L, $\mathrm{u}^{\mathrm{opt}} \rightarrow L^{\mathrm{opt}}$, u being the vector of decision variables defined by the set of indexes df of the input vector, as shown in Equation (1). Common criteria used to select df are the easiness of variable manipulation in the plant, the requirements of industry and the effects of the decision variables on process performance (Basak et al., 2002). The optimisation problem is summarised in Equations (1) and (2):

$$\mathrm{u} = I(\mathrm{df}),\ \mathrm{df} \subset \left\{1, 2, \ldots, \dim_1(I)\right\} \tag{1}$$

where $\dim_n$ refers to the length of the nth-dimension of an array:

$$\begin{aligned} &\hat{\mathrm{u}} = \max_{\mathrm{u}}\ L(\mathrm{I,O}) \\ &\text{s.t.}\quad f(\mathrm{I,O}) = 0 \\ &\mathrm{g(I,O)} \leq 0 \end{aligned} \tag{2}$$

It should be noted that the system evolves with time, although time is not explicitly represented in Equation (2). When steady-state models are considered, process evolution can be represented as a sequence of successive steady-state points. In this text, an element in this sequence is represented either as an array $[\mathrm{I}(\bullet, k)]$ or with the help of a subscript (I_k). It is expected that some elements of I (for instance, feed quality, heat transfer coefficients, environmental temperature and decision variables, among others) change along an operational scenario, being represented by the set of indexes var. Some other elements of I are supposed to remain constant (feed flow and tube diameters, among others), being represented by the set of indexes (std), as shown in Equations (3)–(5):

$$\mathrm{I} = \{\mathrm{I(std)},\ \mathrm{I(var)}\} \tag{3}$$

$$\mathrm{var} \subset \{1, 2, .., \dim(\mathrm{I})\},\ \mathrm{std} = \{1, 2, .., \dim(\mathrm{I})\} - \mathrm{var} \tag{4}$$

$$\mathrm{I(std}, k) = \mathrm{I(std}, 0) = \mathrm{I}_0(\mathrm{std}) \tag{5}$$

COMMON VULNERABILITIES

Incomplete Information

In this text, the information carried by a process is represented by the whole set of process variables $\mathbf{Z} = [\mathrm{I}^T \ \mathrm{O}^T]^T$. Unfortunately, in any real industrial case, the full vector $\mathbf{Z}$ is not available from the process. Process information is primarily obtained through sensors and analysers, which translate physical and chemical properties of streams and equipment into more useful process values. Lack of information is related mainly to the absence of measurements, due to management decisions taken during process design. These decisions are based on sensor costs and known limitations of sensor technology. They also reflect the lack of knowledge about the variables that constitute the real vector $\mathbf{Z}$. As a consequence, the real system is known only through the elements ms of $\mathbf{Z}$, as shown in Equation (6):

$$\begin{aligned} \mathbf{Z} &= \{Z(\mathrm{ms}),\ Z(\mathrm{um})\}, \\ \mathrm{ms} &\subset \{1, 2, \ldots, \dim(Z)\}, \\ \mathrm{um} &= \{1, 2, \ldots, \dim(Z)\} - \mathrm{ms} \end{aligned} \tag{6}$$

If, at instant k, the real plant is under the influence of $\mathrm{I}_k \neq \mathrm{I}_0$, the problem posed to the RTO system may be described in the following terms: starting from the available information set $\mathrm{Q_a} = \{\mathrm{Z(ms},k), \mathrm{I_{0a}}\}$, the RTO has to find out $\mathrm{u}^{\mathrm{opt}} = \mathrm{I}^{\mathrm{opt}}(\mathrm{df})$ that drives the process to $\mathrm{L}^{\mathrm{opt}}$. Therefore, it seems reasonable to ask the following question: how could one predict the "best" state of a process driven by $\mathrm{I}(k)$ if all one knows is some incomplete information of the current state, $\mathrm{Z(ms},k)$, and possibly some additional *available* information at initial state, $\mathrm{I_{0a}}$?

As one can see, the absence of information imposes a severe burden to a RTO system. In order to overcome the incomplete knowledge, a new expanded set of information, Q_{a}^{+}, has to be produced from Q_{a} in such a way that the optimisation procedure can be able to identify the right set $\mathrm{u}^{\mathrm{opt}}$, as shown in Equation (7). This strategy assumes that, in the face of the structure of the process model (f and g), the set of fresh measurements, $\mathrm{Z(ms},k)$, carries an excess of information that can be used to update some elements of the vector of offsets, $\boldsymbol{\Theta}$, which are modifiers of $\mathbf{Z}$, as defined in Equation (8). Due to limited redundancy of the available information, in most problems only a subset of $\boldsymbol{\Theta}$, represented by the index upd in Equation (10), can accommodate the existing excess of information present in the measurements. The remaining set of offsets, represented by the set fix, is supposed to be kept at the (assumed) base values, as shown in Equation (10):

$$Q_{\mathrm{a}} \rightarrow Q_{\mathrm{a}} \rightarrow \mathbf{u}^{\mathrm{opt}} \rightarrow L^{\mathrm{opt}} \tag{7}$$

$$\mathbf{Z}^{+} = \mathbf{Z} + \boldsymbol{\Theta} \tag{8}$$

$$\begin{aligned} \boldsymbol{\Theta} &= \{\boldsymbol{\Theta}(\mathrm{fix}),\ \boldsymbol{\Theta}(\mathrm{upd})\}, \\ \mathrm{upd} &\subset \{1, \ldots, \dim(\boldsymbol{\Theta})\},\ \mathrm{fix} = \{1, \ldots, \dim(\boldsymbol{\Theta})\} - \mathrm{upd} \end{aligned} \tag{9}$$

$$\boldsymbol{\Theta}_k(\mathrm{fix}) = \boldsymbol{\Theta}_{0\mathrm{a}}(\mathrm{fix}) \quad \forall k \tag{10}$$

There are at least two major obstacles that make the accomplishment of the task defined in Equation (7) not simple. First of all, the common implicit assumption that it is possible to correct all uncertain parameters along continuous operation is far from true (Datskov et al., 2006). It is very unlikely that all inputs that vary along the operation scenario are represented by the set of measured variables and updated parameters (i.e. $\mathrm{var} \not\subset (\mathrm{ms} \cup \mathrm{upd})$). There is also a high probability that the number of changing inputs is bigger than the dimension of the set of new information (i.e. $\dim(\mathrm{var}) > \dim(\mathrm{ms}) + \dim(\mathrm{upd})$). The successful implementation of the strategy defined in Equation (7) strongly relies on good instrumentation design (correct and conservative choice of the set ms) and proper formulation of the RTO structure (correct selection of upd) in order to provide enough information and guarantee the reliable estimation of $\boldsymbol{\Theta}(\mathrm{upd})$.

The second obstacle is related to the quality of the initial process information. It is very possible that several variables are assigned to wrong default values at start-up, including thermodynamic and kinetic constants, valve flow coefficients, compositions, environmental conditions, geometry (thickness, length), metallurgical properties of pipes and equipments and so forth. This information is part of I_0, but it is a too strong assumption to expect that it is the same information carried by assumed design values $\mathrm{I_{0a}}$ (Volin and Ostrovskii, 2005). If the wrong default values belong either to the set of unmeasured stimuli, s(um), or to the set of fixed parameters, $\boldsymbol{\Theta}(\mathrm{fix})$, they will be kept wrong forever, leading to inadequate definition of the decision variable values, as shown in Equation (10).

Faulty Information Processing (Model Mismatch)

Another relevant question regarding RTO performance is related to the consequences of model mismatch, that is, the fact that one is unable to find the correct plant structure f and g defined in Equation (2). Since it is very difficult to know the *true* model structure, it is important to understand what features a simplified model should possess in order to allow the optimisation process to find the same optimum of the real plant.

Model adequacy has to be based on optimality conditions, as expressed by the well known Karush–Kuhn–Tucker (KKT) conditions (Fletcher, 1987). Any model must satisfy the KKT conditions at the same set of decision variable values of the real plant in order to present the same $\mathrm{L}^{\mathrm{opt}}$ of the plant. Although dealing with a slightly different problem, Biegler et al. (1985) used this concept to present some conditions of optimality that must be satisfied by RTO systems. According to Biegler et al. (1985), a simplified model can be regarded as appropriate (and lead to RTO responses that are similar to results obtained with a more rigorous model) if the gradients calculated with the simplified model match the gradients of the rigourous model at all points. This means that $\nabla_{\xi}\, \mathbf{g}_{\mathrm{m}} = \nabla_{\xi}\, \mathbf{g}_{\mathrm{p}},\ \nabla_{\xi} f_{\mathrm{m}} = \nabla_{\xi} f_{\mathrm{p}},\ \forall \xi = [\mathrm{I}^{\mathrm{T}}\, \mathrm{O}^{\mathrm{T}}]^{\mathrm{T}}$, where subscripts p and m indicate the plant (or the rigorous model) and the simplified model respectively. However, as the researchers acknowledge, this implies that only the rigourous model can be eventually regarded as appropriate for optimisation and that the RTO scheme based on the simplified model is not guaranteed to provide the real optimum operation point of the plant. Similar conditions were also presented by Forbes et al. (1994) in terms of a less severe pointwise model adequacy criterion by assuming the stationarity of the gradient, the negative definiteness of the Hessian of active constraints and the reduced gradient equations in the reduced space of the optimisation problem. From a pragmatic point of view, it can be concluded that a simplified model cannot replace a more rigourous model in rigourous mathematical terms.

A real-time optimiser faces a difficult challenge that can be defined as: how should the imperfect model be adapted to an ever-changing process in order to always find the best set of decision variables? Most RTO implementations do not characterise the model adequacy explicitly and quantitatively when input vector I follows a generic path in the space $R^{\dim(I)}$, and some sort of adaption has to be performed in presence of incomplete

information. Besides, a fundamental problem persists: how can one know what is changing when the set of information is not complete? How can one know unequivocally if the elements of **I** (and not the model structure *f*) are changing? Unfortunately, these two questions remain unanswered, which means that the validity of local adaption procedures for non-local sets of model predictions is questionable on fundamental mathematical grounds.

There are at least three possible consequences of the fact that the model structure cannot be perfectly adapted in the analysed operational region: (i) selected decision variables will lead to suboptimal operation conditions; (ii) some constraints in **g** will possibly be violated and (iii) more than one RTO cycle will be necessary to stabilise the set of predicted decision variables. This last consequence is related to the fact that RTO optimisation assumes that the process operation will produce outputs $\hat{\mathbf{O}}$ after implementation of the set of decision variables, $\hat{\mathbf{u}} \rightarrow \mathbf{u}$. If $\hat{\mathbf{u}} \rightarrow \hat{\mathbf{O}}$ does not materialise, then the RTO procedure will explain this fact by means of a new set of adjusted parameters, which will lead to another set of decision variables, as summarised in Equation (11):

$$\text{If } \mathbf{u}_k \rightarrow \mathbf{O}_{k,\text{plant}} \neq \hat{\mathbf{O}}_{k,\text{ mod el}} \Rightarrow \boldsymbol{\Theta}_k(\mathbf{upd}) \neq \boldsymbol{\Theta}_{k-1}(\mathbf{upd}),\ \mathbf{u}_{k+1} \neq \mathbf{u}_k \tag{11}$$

Although it may take several RTO cycles until final process stabilisation can be attained, there is no guarantee that process will be at optimal condition at the end. These features can be seen through a simple example, as proposed originally by Biegler et al. (1985) and Zhang and Forbes (2000) and shown in Equations (12–14):

$$\min_x\ L = (y - 1/2)^2 + (x - 1)^2 \quad \text{s.t.}\quad f(x, y) = 0 \tag{12}$$

$$\text{Plant}: \lfloor (x-1)^3 + (x-1)^2 + 1 \rfloor - y = 0 \tag{13}$$

$$\text{Model}: x + \beta - y = 0 \tag{14}$$

According to the proposed formulation:

$$\mathbf{I} = [x \quad \beta]^{\mathrm{T}}, \quad \mathbf{O} = y$$

$$\mathbf{df}=1; \quad \mathbf{u} = x$$

$$\mathbf{Z}^{+} = \mathbf{Z} + \boldsymbol{\Theta} = [x\ \beta\ y]^{\mathrm{T}} + [\delta x\ \delta\beta\ \delta y]^{\mathrm{T}}$$

$$\mathbf{upd} = 2$$

$$\mathbf{fix} = [1\ 3]$$

$$\boldsymbol{\theta} = \boldsymbol{\Theta}(\mathbf{upd}) = \beta$$

$$\boldsymbol{\Theta}_0 = [0\ 0\ 0]^{\mathrm{T}}$$

As shown in Figure 1, it takes about 8 RTO cycles to stabilise the decision variable *x*. This can constitute a very long period of time, depending on the particular RTO interval. For instance, if the RTO interval is equal to 1 h in a typical continuous polyolefin process, the attainment of the local optimum can be useless for practical purposes, since the product lot can be finished in <8 h. Similarly, if the RTO interval is equal to 2 h, as in many refineries, the attainment of the local optimum can be useless because of the several disturbances, such as change in feed composition, may occur along the 16 h of stabilisation time. Even worse, after the stabilisation period, the process is left at a suboptimal point. Of greater concern in Figure 1 is the fact that the process was placed at the optimum point at the beginning of the RTO cycle, but it was forced to move apart from the optimum point because of RTO intervention (although the expected performance should be exactly the opposite). In many problems, model parameters are related to some concrete real world meaning; however, if model mismatch is present, updated parameters can be meaningless, as observed in Figure 1 for parameter β.

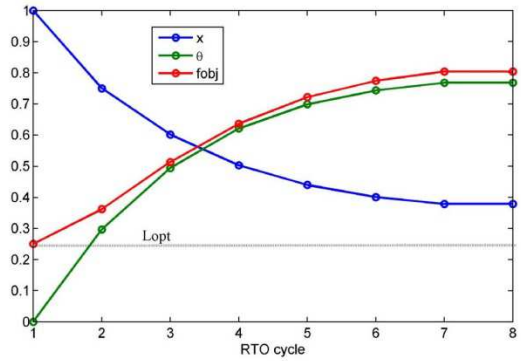


Figure 1. RTO implementations of optimised decision variable, *x*, in presence of model mismatch. β: updated parameter; *L*: objective function of top optimisation layer.

Corrupted Information

Information provided by sensors is expected to be similar but not equal to the "true" information produced by the process. Sensors incorporate into the process signals some features that are not related to the behaviour of the "true" variables. This creates another challenge to RTO implementations, since the RTO scheme is expected to produce reliable values for decision variables, although the whole optimisation process is based on real (and possibly unreliable) signals within an incomplete set of information.

The first thing one should know about measured data is the nature of signal corruption, if one intends to be ready to cope with its consequences. It is usually assumed that the signal that is used by the RTO has two components: a deterministic component (the "truth") and a stochastic component (the "noise"), which are combined additively according to Equation (15):

$$\mathbf{Z}_{\text{meas}} = \mathbf{Z}_{\text{true}} + \varepsilon \tag{15}$$

Elements of $\boldsymbol{\varepsilon}$ are assumed to be sampled from a population of known probability density function (p.d.f.), $\psi(\boldsymbol{\varepsilon})$. The p.d.f. is commonly assumed to be a Gaussian function (Roberts and Williams, 1981; Forbes and Marlin, 1996), although this does not correspond to reality most often.

It is convenient to define $\mathbf{z} = \mathbf{Z}(\mathrm{ms})$ from Equation (6) as the set of all measured variables. In the context of the two-step RTO scheme, the adaptation step performs the task of selecting the set $\theta = \Theta(\mathrm{upd})$ that better explains measurements $\mathbf{z}$ in the light of the plant model. In order to do that, besides the model structure, it is also necessary to take into account the probability of occurrence of noise. In other words, it is necessary to answer the realistic question: in face of the real corrupted measurements $\mathbf{z}$, what is the set θ that *most likely* gives rise to $\mathbf{z}$? This question constitutes the statistical cornerstone of maximum likelihood adaptation schemes (Bard, 1974), as described in Equation (16), which consists of maximising the likelihood of function F under constraints imposed by the process model. It should be noted that the elements of $\mathbf{z}$ included in the objective function are those related to the indexes obj, where $\dim(\mathrm{obj}) \leq \dim(\mathbf{z})$, according to decisions taken during the design of the RTO structure.

$$\begin{aligned} &\hat{\theta} = \arg\max_{\theta} \; F(\mathbf{z}(\mathrm{obj})) \\ &\text{s.t.} \quad f(\mathrm{I},\mathrm{O}) = 0 \\ &g(\mathrm{I},\mathrm{O}) \leq 0 \\ &\theta = \Theta(\mathrm{upd}) \\ &\mathrm{obj} \subset \{1, \ldots, \dim(\mathbf{z})\} \end{aligned} \tag{16}$$

If $\boldsymbol{\varepsilon}$ follows a Gaussian p.d.f., the problem defined in Equation (16) becomes the weighted least squares (WLS) estimation:

$$\begin{aligned} \hat{\theta} = \arg\max_{\theta} \; & (\mathbf{z}_{\mathrm{meas}}(\mathrm{obj}) - \mathbf{z}_{\mathrm{model}}(\mathrm{obj}))^{\mathrm{T}} \, \mathbf{V}_z^{-1} (\mathbf{z}_{\mathrm{meas}}(\mathrm{obj}) \\ & - \mathbf{z}_{\mathrm{model}}(\mathrm{obj})) \\ \text{s.t.} \quad & \mathrm{f}(\mathrm{I},\mathrm{O}) = 0 \\ & g(\mathrm{I},\mathrm{O}) \leq 0 \end{aligned} \tag{17}$$

where $\mathbf{z}_{\mathrm{meas}}$ represents values of $\mathbf{z}$ acquired with the help of process instrumentation and $\mathbf{z}_{\mathrm{model}}$ are values of $\mathbf{z}$ manipulated by θ according to the model structure (f, g). $\mathbf{V}_{\mathbf{z}}$ is the covariance matrix of $\mathbf{z}_{\mathrm{meas}}$, which is normally assumed to be diagonal (measurement fluctuations are assumed to be independent, which very often cannot be supported by sound statistical analysis of available process data).

However, several fundamental assumptions are behind the formulation presented in Equation (17) (Bard, 1974; Bates and Watts, 1988): (i) noise $\boldsymbol{\varepsilon}$ is normally distributed with zero mean; (ii) input variables I are known perfectly and are free of noise and (iii) $\boldsymbol{\varepsilon}$ elements are distributed independently.

If the third assumption is not valid, the full covariance matrix $\mathbf{V}_{\mathbf{z}}$ should be used in order to assure that the estimator is unbiased and efficient (Bard, 1974; Bates and Watts, 1988). However, the full covariance matrix of measurement noise is almost never characterised at plant site, meaning that real estimators implemented at plant site are likely to be biased and inefficient, even if measurement noises follow the Gaussian function, which is certainly doubtful. If the previous assumptions are valid and if the process model is linear, in the absence of active constraints, it can be proved (Bates and Watts, 1988) that $\psi(\hat{\theta})$ is normally distributed and that the confidence region of the estimated parameters is a continuous hyper-ellipsoid (Seber and Wild, 1989).

Despite the fact that significant deviations from the previous assumptions can be found in virtually all real world industrial cases, validity of the mentioned assumptions is almost never checked. For instance, common commercial RTO software does not provide tools for estimation of the covariance matrix $\mathbf{V}_{\mathbf{z}}$, making the existence of independently distributed $\boldsymbol{\varepsilon}$ a non-verifiable premise. Besides, evidences on non-Gaussian distributions of measurement noises are commonly interpreted in practice as the existence of gross-errors, although such evidences should be used to support the pursuit of the real nature of the $\boldsymbol{\varepsilon}$ signal. As a matter of fact, the use of Equation (17) without statistical validation of the underlying assumptions put in risk valuable properties of the maximum likelihood estimators, such as consistency, asymptotical normality and efficiency. Relying on the Central Limit Theorem (Bard, 1974; Bates and Watts, 1988) to justify the use of normal distributions and assuming absence of calibration drifts to validate the zero mean hypothesis may be too optimistic.

Table 1. Noise parameterisation

Cases	ϕ	σ_x	σ_y
A	0	0	0.05
B	0.3	0.048	0.048

The following example illustrates how parameter estimation may produce misleading results when underlying hypotheses are not satisfied. Let us consider a real world process that produces output signal y as function of x, α and β, according to Equation (18). Input variables α and β are assumed to vary along time and are estimated via the respective offset corrections. The set of measured variables is $\mathbf{z} = [\mathrm{x}\ \mathrm{y}]^T$ and only y is explicitly included in the formulation of the objective function (index vector $\mathrm{obj} = 2$). $\mathbf{V}_z$ is defined as the identity matrix, thus ignoring the auto-correlated nature of $\boldsymbol{\varepsilon}$.

Process information comes through measurements corrupted by additive auto-correlated noise, as shown in Equation (19). Independent variable $\mathbf{x}$ is linearly spaced in the range [0 5]. Model predictions follow Equation (20). Results obtained with two different noise configurations (see Table 1) were compared to each other. In Case A, all assumptions behind Equation (17) are valid; on the other hand, in Case B, input signal is not known perfectly and noise is auto-correlated. It should be noted that σ_y is slightly different from Case A to Case B in order to produce the same standard deviation of ε for both cases.

Real world:

$$\mathrm{y} = 0.3\alpha + 2.1\beta \frac{\mathrm{x}}{0.4 + \alpha\mathrm{x}} + 0.4\beta\mathrm{x} \tag{18}$$

Measurements:

$$\begin{aligned} &\mathrm{y}_{\mathrm{meas}} = \mathrm{y} + \boldsymbol{\varepsilon}_y; \;\; \mathrm{x}_{\mathrm{meas}} = \mathrm{x} + \boldsymbol{\varepsilon}_x \\ &\varepsilon(n) = \phi\varepsilon(n-1) + \varphi(n); \quad \varphi\tilde{N}(0, \; \sigma) \end{aligned} \tag{19}$$

$$\mathrm{y}_{\mathrm{model}} = 0.3\alpha + 2.1\beta \frac{\mathrm{x}_{\mathrm{meas}}}{0.4 + \alpha\mathrm{x}_{\mathrm{meas}}} + 0.4\beta.\mathrm{x}_{\mathrm{meas}} \tag{20}$$

In order to compare the results obtained with the two different noise structures, 20 000 replications of the estimation procedure were performed for each case. In the present study, the "true" parameter values are $\alpha = 2.9$ and $\beta = 1.7$. Since the process model is non-linear in the parameters, the confidence regions of parameter estimates were not expected to present a perfect ellipsoidal shape, as confirmed in Figure 2. It is interesting to note the significant deviations from the ellipsoidal confidence region when the underlying statistical assumptions are violated (Case B), as described by Schwaab et al. (2008). However, the worst

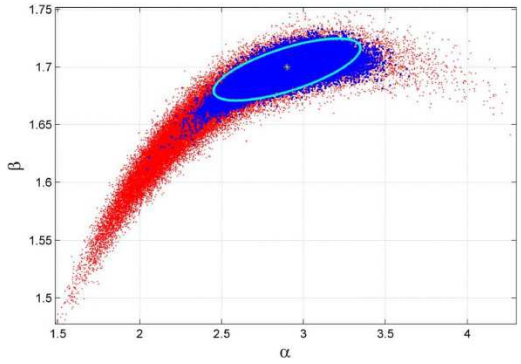


Figure 2. Estimated parameters for 20 000 replications of Case A (blue), Case B (red). Cyan ellipse surrounds 95% confidence region for linearised model. Yellow star indicates true (α, β) pair.

consequence of Case B is that there is no asymptotic guarantee that unbiased estimated parameters are expected to be produced. In the present example, the mean percentage deviation observed between the estimated values and the true value was equal to -11.4% for α and -1.6% for β. In the context of the two-step RTO scheme, propagation of structural imperfection (wrong parameter values) resulting from the estimation procedure to the upper optimisation level will cause suboptimal selection of decision variables and increase the chance of constraint violation.

Another relevant issue regarding the parameter estimation step is related to the estimability of the updated parameters. The confidence region of the estimated parameters can become very large and even unbounded (Schwaab et al., 2008), due to factors such as the choice of $\boldsymbol{\Theta}(\mathbf{upd})$, the plant operating conditions and the data set used (Miletic and Marlin, 1998). Identification of estimability problems can constitute a major problem at plant site for a number of reasons, including: (i) the model structure can change along the time; (ii) estimability analysis depends on the available data, which change along the RTO cycles; (iii) the sensitivity analysis required to build the Hessian matrix of the objective function and formally characterise the rank of the estimation problem can constitute a numerical task that is too expensive to be performed iteratively at plant site and (iv) rigorous estimability analysis depends on the unknown parameter estimates. For all these reasons, estimability problems are frequent and should be expected during the implementation and use of RTO schemes. Despite that, available commercial software does not address this fundamental issue of the parameter estimation step on firm mathematical grounds. As shown in the next sections, however, estimability analyses can be of fundamental importance to assure the proper performance of the RTO scheme at plant site.

The quality of process measurements can be significantly improved if the parameter estimation algorithm is combined with data reconciliation procedures and gross error detection in order to reduce noise variability (Prata et al., 2009). Basically, a data reconciliation problem is a parameter estimation problem that includes the manipulation of measured variables in order to force available data to be in accordance with process model structure. In formal grounds, the so-called data reconciliation procedure occurs when $\mathbf{ms} \cap \mathbf{upd} \neq \phi$. As a consequence, data reconciliation procedures can produce parameter estimation problems of very high dimension. Therefore, numerical problems associated with the distinct sensitivity of the objective function in respect to the different process variables can be expected when data reconciliation procedures are implemented in real problems. In these cases, the detailed estimability analysis of the proposed problem can be of fundamental importance, as the inputs can exert effects of very different magnitudes on the objective function. In most scenarios, only the most influential variables can be estimated with the available data, while the remaining variables must be kept constant their nominal values. The estimation problem must be carefully designed if proper performance is expected at plant site.

Miletic and Marlin (1998) proposed the singular value decomposition (SVD) of the Fisher information matrix for design of the estimation problem. The Fisher information matrix can be roughly defined as the Hessian matrix of the objective function. The parameters selected to define the estimation problem (including inputs, in data reconciliation procedures) must necessarily guarantee that the Fisher information matrix is nonsingular (and invertible). In more practical terms, it can be said that the characteristic values of the Fisher information matrix must be above a certain critical limit, which depends on the magnitudes of the analysed variables. Similar techniques have also been used with empirical principal component analysis (PCA) models and selection of parameters in kinetic problems (McLean and McAuley, 2012).

Numerical Optimisation

The two-step RTO scheme is based on successive nonlinear optimisation problems, which are subject to equality and inequality constraints. As previously discussed, several problems can jeopardise the success of the RTO task. Unfortunately, some reasons for concern still remain even when the complete and perfect set of information is available and the model structure is known exactly. Previous discussion relies on the implicit assumption that there is a numerical method able to find the right set of manipulated variables in both optimisation problems (2) and (17). However, features of the objective function, such as the curvature, the existence of discontinuities and the possible existence of multiple relative maxima/minima, may pose very challenging problems to numerical optimisation algorithms.

Deterministic methods of optimisation, in particular sequential quadratic programming (SQP), are largely employed in RTO systems, both in the academic literature and in commercial software. Although SQP methods have well-known limitations (Yang, 2010), the RTO literature does not report the impact of numerical limitations on the performance of the RTO scheme very frequently (perhaps because unsuccessful implementations are not discussed very often in the technical literature). Few articles (Cubillos et al., 2007; Golshan et al., 2008) clearly mention flaws related to such optimisation methods. Particularly, Golshan et al. (2008) showed how SQP schemes tend to get trapped by local minima and fail to converge when intermediate calculations fall in the infeasible region of decision variables. Prata et al. (2009) also showed that the rate of failure of SQP schemes could be unacceptable when dynamic data reconciliation problems were solved in real time in a polymerisation process.

The process economic performance will certainly be degraded when failure to attain the true optimum points takes place. Besides, the two-step RTO scheme will be prone to produce wrong sets of parameter estimates and decision variables, leading to poor process operation. A simple example, as described in Equation (21), is helpful to show how performance degradation can be related to the numerical optimisation method, available initial

Table 2. Deviation from optimal **u** as a function of initial guesses and optimisation method

$\Delta u\%$	$[-2\ -2]$	$[2\ 2]$
NM Simplex	94.5	35.6
SQP	88.0	35.6

guesses and the specific shape of **f**. In this case, all information regarding x_1, x_2 and y is perfectly known and no parameter is estimated. Performance is measured in terms of relative deviation from the optimum set of decision variables and from the optimum value of the objective function, as defined in Equation (22). As one can see in Tables 2 and 3, optimised decision variables and the objective function can be affected differently by the available initial guesses. Results from two optimisation methods are presented: SQP and Nelder Mead's Simplex (Nelder and Mead, 1965), referred as NM simplex. While all optimisation methods (wrongly) agree about what must be done when the initial guesses are equal to [2 2], they indicate different movements when the initial guesses are equal to $[-2\ -2]$ and lead to suboptimal values of L. A major problem, though, is that it is not possible to know in advance how biased optimisation results will be. Unfortunately, there is no guarantee that a certain optimisation method will perform better than others, since it can be proved (Wolpert and Macready, 1997; Koppen, 2004) that there is no optimisation algorithm able to perform better than any other for any class of problems.

$$\begin{aligned}&\hat{\mathbf{u}} = \arg\max_{\mathbf{u}}\ L\\ &\text{s.t.}\quad {x_1}^2 + (x_2 - 1.2)^2 - x_1x_2 + x_1\sin(2\pi x_2)\\ &\qquad + x_2\sin(2\pi(x_1 + 0.5)) - y = 0\\ &\text{where}\quad u = [x_1\ x_2]^{\mathrm{T}},\quad L = y\end{aligned} \tag{21}$$

$$\Delta\mathbf{u}\% = \left\| 100\frac{\hat{\mathbf{u}} - \mathbf{u}^{\mathrm{opt}}}{\mathbf{u}^{\mathrm{opt}}} \right\|, \qquad \Delta L\% = 100\frac{\widehat{L} - L^{\mathrm{opt}}}{L^{\mathrm{opt}}} \tag{22}$$

It can be learned from Tables 2 and 3 that an imperfect optimisation method may impact the process performance, as described in Equation (22), in ways that are hard to predict. Depending on the initial guesses, the optimisation method may drive the solution to different regions due to the influence of the vicinity of different local extrema. In the absence of corrupted information, given a certain continuous p.d.f. of the initial guess $\mathbf{u}_0$, a RTO using an ideal deterministic optimisation method would produce a dirac delta function for $\psi(L)$. However, non-ideal methods may possibly produce a different p.d.f. $\psi(L)$, as can be seen in Figure 3, where the p.d.f. $\psi(L)$ is in a scenario of uniformly distributed initial guesses, $\mathbf{u}_0$, over the region $|u_0(1)| \leq 2$, $|u_0(2)| \leq 2$.

Stochastic methods can be used to improve the robustness of the optimisation. Although the use of these methods tends to

Table 3. Deviation from optimal L as a function of initial guesses and optimisation method

$\Delta L\%$	$[-2\ -2]$	$[2\ 2]$
NM Simplex	43.4	29.7
SQP	62.1	29.7

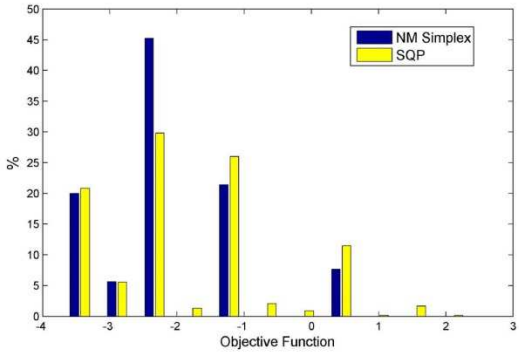


Figure 3. Probability density function of optimised objective function when initial guess is uniformly distributed over the region $|u_0(1)| \leq 2$, $|u_0(2)| \leq 2$.

increase the computational time, the asymptotically global character of the optimisation can be guaranteed. It must be noticed that the stochastic nature of the optimisation method leads to additional changes in the shape of p.d.f.s defined in Equation (23). In spite of that, deviations from optimum values tend to be lower because of the enhanced exploration of the feasible region. Figure 4 shows some results obtained for the problem defined in Equation (21) when it was solved with two different stochastic methods: Particle Swarm Optimisation (PSO) and Genetic Algorithms (GA). These methods evolve from an initial population (set of initial guesses) in successive steps (generations). The population of guesses changes along the generations through combination, randomisation and selection of individual characteristics.

In Figure 4, each point is obtained as the result of optimisations performed with different population sizes and number of generations. The results are expressed in terms of the 90th percentile of $\Delta L\%$ and the 90th percentile of the calculation time. It is important to say that all but one element of the initial population was selected at random. The deterministic element was the process state before the optimisation. Some noticeable facts presented in Figure 4 are: (i) performance measurements are influenced significantly by parameterisation; (ii) optimisation results are stochastic even when all inputs are deterministic ($\boldsymbol{\varepsilon} = \mathbf{0}$); (iii) the initial population affects the results and (iv) it is not possible to name the best optimisation method. The performance of the optimisation method depends on the initial population, the accepted tolerance for sub-optimality and the importance of the computation time. As one can see in Figure 4, while PSO is clearly superior to GA when $[-2\ 2]$ is an element of the initial population, this result is far less clear when the initial population contains [2 2] as a deterministic element.

As discussed in the previous section, if some input to RTO is stochastic in nature, all results will share this feature. It means that one can see the RTO scheme as a system that creates sets of probability density functions in several spaces. All these p.d.f.s are generated from the single p.d.f. of noise $\boldsymbol{\varepsilon}$ (23):

$$\psi(\boldsymbol{\varepsilon}) \xrightarrow{\text{RTO}} \psi(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{RTO}} \left\{\psi(\hat{\mathbf{u}}),\ \psi(L)\right\} \tag{23}$$

In order to estimate the p.d.f.s described in Equation (23), some very popular assumptions are used, including the assumption that

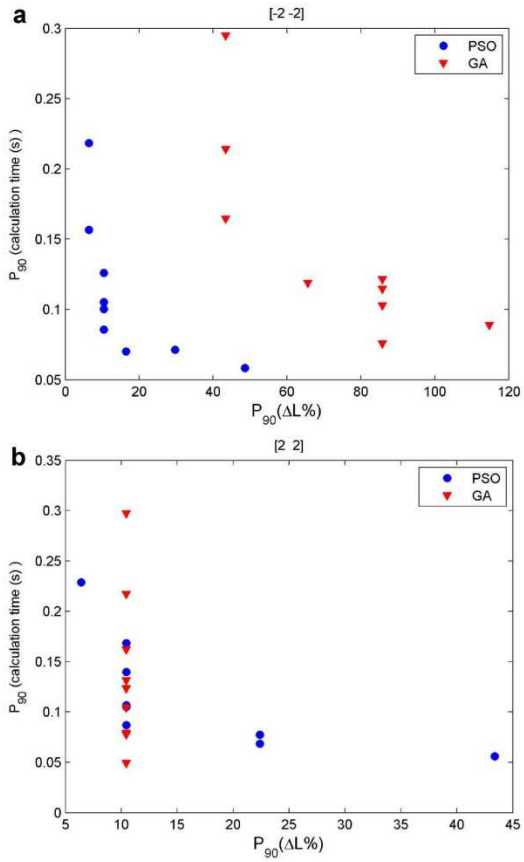


Figure 4. a: 90th percentiles of $\Delta L\%$ and calculation time for optimisation algorithms PSO and GA under different parameterisations (population size and number of generations) when point $[-2\ -2]$ is part of the initial population. b: 90th percentiles of $\Delta L\%$ and calculation time for optimisation algorithms PSO and GA under different parameterisations (population size and number of generations) when point [2 2] is part of the initial population.

ε follows a Gaussian distribution and that the process model is linear (Miletic and Marlin, 1996, 1998). If these assumptions are valid, $\psi(\hat{\theta})$, $\psi(\hat{\mathbf{u}})$ and $\psi(L)$ are also Gaussian, and the confidence regions are continuous hyper-ellipsoids (Seber and Wild, 1989) in the space of the related variables. The fact that the RTO structure integrates two successive optimisation layers, paying attention to very different non-linear objective functions, may lead to probability density functions that are very different from Gaussian distributions in real cases.

A less obvious consequence from the present discussion is that the p.d.f.s sequentially created by the RTO, as indicated in Equation (23), are profoundly affected by the optimisation method and its parameterisation, even if deterministic algorithms are employed. In a real, large-scale non-linear problem, with changing sets of active constraints, influence of the optimisation method on the process indicated in Equation (23) is still more pronounced.

CASE STUDIES

Depropaniser

A propene/propane splitter (distillation column with 165 equilibrium stages) is used to illustrate some vulnerabilities of the standard two-step RTO scheme discussed in the previous sections. The feed stream contains propane, propene and butane with nominal relative composition of (38:57:5) in molar basis. The condenser is a standard shell/tube heat exchanger, and the cold fluid (water) flowrate is kept constant at the maximum allowed value. Steam is used in the reboiler, as usual, and reboiler simulations assume that the energy input is known. The RTO task comprises the profit maximisation by using the available degrees of freedom: the feed flowrate (Q_{feed}), the reflux ratio (RR) and the reboiler (H_{reb}) heat duty (power). Molar fraction of propene ($x_{\text{pe,tp}}$) at the distillate stream cannot be lower than 0.994.

Table 4 shows the nominal values and the upper and lower bounds for decision variables. The economic objective function (profit) is the difference between propene sales revenue and the cost of heat power delivered to the reboiler (Equation 24). It should be noted that if the reboiler power is greater than a certain critical value, H^0_{reb}, an extra source of energy is used ($\$_{H2} > \$_{H1}$). Simulations were performed with a proprietary process simulator.

$$L = Q_{\text{tp}} x_{\text{pe,tp}} \$_{\text{pe}} - \min(H_{\text{reb}},\ H^0_{\text{reb}})\$_{H1} - \max(0,\ H_{\text{reb}} - H^0_{\text{reb}})\$_{H2} \tag{24}$$

In the presence of complete and non-corrupted information, the optimal set of decision variables should be selected by the RTO procedure; however, as already discussed, perfect optimisation is not available. The first example was designed to show the deviation from optimum value as a function of the selected optimisation methods. In order to illustrate this dependency, it is assumed that the model is perfect, that the input set I is constant and that there is no significant noise measurement. By doing so, it is possible to skip the parameter estimation step and focus specifically on the top optimisation layer of RTO.

It can be seen from Table 5 that, for all methods, the set of optimised decision variables led to suboptimal values of L. As already discussed, parameterisation of the optimisation methods can strongly influence the obtained results. Definition of the best optimisation method is not pursued here for two reasons: (i) there is no such method for the general case (Wolpert and Macready, 1997; Koppen, 2004) and (ii) the main objective is to show that common problems may pose difficulties to standard RTO algorithms. In order to obtain at least a "fair" set of values, care was taken to ensure that optimisation was stopped when both the objective function and decision variable values were smaller than the specified tolerances (relative deviations of 10^{-4}). Tolerances were kept the same in all simulation studies.

Table 4. Range for decision variables in the depropaniser example

	Decision variables		
	Q_{feed} (kmol/h)	RR	H_{reb} (kW)
Lower	400	0.92	5000
Nominal	446	0.94	14 500
Upper	500	0.96	21 894

Table 5. Optimised decision variables and profit values for different deterministic optimisation methods

	Q_{feed} (kmol/h)	RR	H_{reb} (kW)	L
Start point	446	0.940	14 500	155.7
Optimum	464	0.930	13 404	180.0
NM Simplex	464	0.933	13 857	178.4
Direct	420	0.936	14 000	163.3
SQP	446	0.937	14 501	165.3

Table 6 shows deviations from optimum according to two different metrics, $\Delta u\%$ and $\Delta L\%$, as defined in Equation (22). It can be seen that the best profit result ($\Delta L\% = -0.9$, from optimum) corresponds to the least deviated set of decision variables ($\Delta u\% = 3.4$, from optimum). However, such correlation should not be regarded as a rule. If the starting point is taken as reference, all methods improved the objective function, although most of them changed decision variables less than optimal $\Delta u\% = 8.5$. The Direct method (Bjorkman and Holmstrom, 1999) was the method that changed **u** (6.8%) most, but the method performed poorly in terms of optimum objective function values. This means that the process would be severely disturbed by the proposed process change, although the proposed movement would not necessarily drive the process to the optimal direction.

Values shown in Tables 5 and 6 for deterministic optimisation methods will always be the same, provided that initial guesses and parameterisation are not altered. On the other hand, results produced by stochastic methods could be interpreted as samples taken from a certain p.d.f. In this sense, $\Delta u\%$ and $\Delta L\%$ are random variables as well as the amount of time spent by the optimisation algorithm (t_{exec}). Figure 5 shows the 90th percentiles for distributions of $\Delta L\%$ and algorithm execution time for three different hybrid methods: Direct-PSO-NM Simplex (Dir-PSO), PSO-NM Simplex (PSO) and GA-NM Simplex (GA). Each percentile value is calculated over 200 results of repetitions of the optimisation procedure performed with the same parameterisation.

The choice of the "best" parameterisation of the "best" hybrid method is dependent on the relative importance of the features shown by each p.d.f. generated, in the sense of Equation (23). For example, the user could choose the best configuration as the one that produces the fastest result [lowest 90th percentile of $\psi(t_{exec})$] within an acceptable degree of suboptimality (90th percentile of $\Delta L\%$ below 2%). Tuning parameters are the population size and the number of generations of the PSO and GA. According to this criterion, the best configuration was obtained with the PSO algorithm, with population of 40 individuals evolving for 5 generations. The 90th percentile for this configuration is emphasised by a green circle in Figure 5a. The random nature of the variables is made explicit in Figure 5b, where the distribution of $\Delta L\%$ for the best configuration is presented.

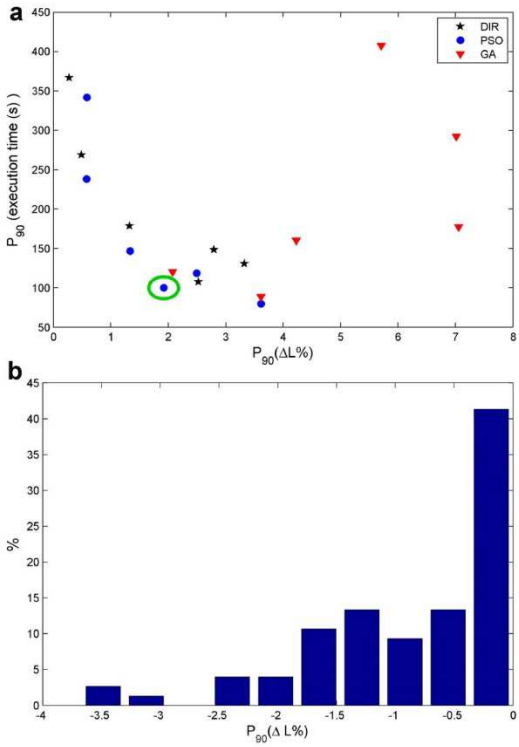


Figure 5. a: 90th percentile of $\Delta L\%$ and 90th percentile of execution time for different hybrid methods. Each point is related to a different parameterisation (population size, number of generations) of three different hybrid optimisation methods. Green circle indicates the best method and parameterisation. b: Histogram of 200 $\Delta L\%$ results for the best configuration.

In order to be more realistic, the simulation scheme was modified in order to accommodate for the unavoidable RTO vulnerabilities: (i) measurements were assumed to be subject to additive Gaussian noise, $\varepsilon \sim N(\mathbf{0}, \Sigma)$; (ii) the covariance matrix Σ was assumed to be diagonal; (iii) the average of each measured process signal over the time window of data acquisition has standard deviation equal to 1% of the nominal value; (iv) parameters UA (heat transfer coefficient in the condenser) and η (column efficiency) were estimated to allow for fitting of available process data and (v) the molar fraction of propene in the distillate stream could not be lower than a specified minimum value.

Table 6. Optimised deviations of decision variables and profit values for different deterministic optimisation methods

	$\Delta u\%$ (from optimum)	$\Delta L\%$ (from optimum)	$\Delta u\%$ (from start)	$\Delta L\%$ (from start)
Start point	9.0	−13.5	0.0	0.0
Global optimum	0.0	0.0	8.5	15.6
NM Simplex	3.4	−0.9	6.1	14.5
Direct	10.2	−9.3	6.8	4.8
SQP	8.9	−8.1	0.3	6.1

Table 7. Estimated parameters: allowed range

	Adjustable parameters	
	UA (kW/C)	Efficiency
Lower	1600	0.8
Nominal	2800	1
Upper	2800	1

Simulation runs consisted of five successive RTO cycles (parameter updating + economic optimisation), where only decision variables produced by the RTO and updated model parameters were allowed to vary (remaining inputs were assumed to be constant and process outputs were updated through simulation and corrupted with the assumed random noise). Series of five cycles were repeated 100 times in order to characterise the distribution of output variables.

The choice of updated parameters was made according to common sense knowledge that RTO should provide plant diagnostics besides plant optimisation. In this sense, plant staff agrees that global heat transfer coefficient times the area of top condenser (UA) and Murphree tray efficiency (η) may be reasonable choices for the set of updated parameters, since it is expected they show some degradation as time passes. Upper and lower bounds for estimated parameter values were fixed based on previous empirical knowledge and are shown in Table 7.

Histograms of all 500 estimated values (5 cycles times 100 repetitions) of each parameter are shown in Figure 6. Although $\boldsymbol{\varepsilon}$ is assumed to follow the normal distribution, one can see that parameter estimates are not Gaussian, as the process model is nonlinear. The shape of Figure 6 is strongly affected by the fact that estimated values are often constrained by the defined bounds, which is usually ignored by conventional statistical analyses. Besides, the η distribution is strongly influenced by the fact that the objective function is relatively insensitive to modification of η values in the proposed problem. As a result, despite the well-defined physical meaning of estimated parameters, physical interpretation of parameter estimates is doubtful.

Since the input set $\mathbf{I}$ is not changed (except for the subset $\mathbf{u}$), changes made in UA and η during parameter updating are responsible for economic optimisation failure. If values of estimated parameters are different from the real values, the second optimisation layer will be, in fact, optimising the wrong process. As a consequence, model mismatch derived from changes of η and UA from their real values, in addition to inherent flaws of the optimisation algorithms, can lead to degradation of the RTO performance.

In the previous case, shown in Figure 5, RTO performance was affected only by imperfections of the optimisation method. In the present case, there is an additional burden represented by the noise vector $\boldsymbol{\varepsilon}$, which increases the probability of sub-optimal optimisation, as one can observe when the histograms of Figures 5 and 7 are compared to each other. While in Figure 5 there was 90% of chance (best configuration) for profit to be 2% below the optimum value, in the second case (Figure 7), deviations from optimum values will be in the range 0–15% with the same probability. Maybe the worst consequence is related to the variability of decision variables. It can be seen from the $\Delta u\%$ histogram in Figure 7 that most of time $\hat{\boldsymbol{u}}$ is far from $\mathbf{u}^{\text{opt}}$. It is also interesting to notice the spread in $\Delta u\%$, which indicates the excessive rate of movements forced by the RTO. This fact is made clearer in Figure 8, which shows the

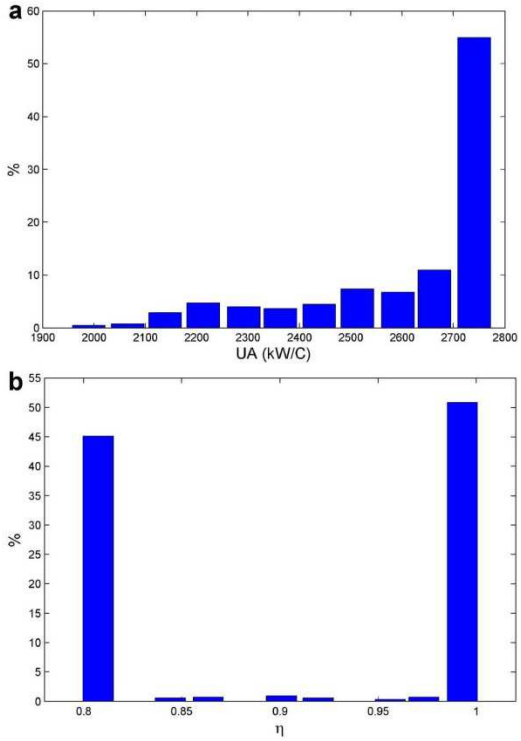


Figure 6. Histogram of estimated parameters in the splitter case. Values were generated through 100 repetitions of five consecutive RTO cycles.

percentage difference, $\mathrm{d}(x) = 100(x_k - x_{k-1})/x_{k-1}$, between decision variable values of two successive RTO cycles. It is a matter of concern to observe that the two-step RTO scheme can change decision variables significantly in the short term.

It is important to notice that many decisions related to the RTO structure may mitigate (or further amplify) some of the observed performance problems. It can be seen in Figure 9 that some changes in the set of updated offset parameters may significantly decrease the variability of decision variables. Under the point of view of $\mathbf{u}$ behavior, the median deviation from optimum (red lines) as well as variability along RTO cycles (width of rectangles and vertical lines) indicate that the choice of (UA,η) as updated offset parameters is probably the worst. Much better choices would be the updating of RR (additive offset to measured RR) or (UA,RR). These choices are also supported by some data from Table 8, as the median of $\Delta L\%$.

Since some model mismatch is artificially created by badly estimated parameter values, it can be expected that process constraint violations will also be impacted by the poorly optimised decision variables, as a consequence of the poorer model performance. Table 8 shows constraint violations (minimal propene molar fraction in the top product) as a function of the analysed estimation problem. The frequency of constraint violations is surprisingly high in some cases, although the optimisation problem does consider the process constraint during the optimisation step. The obtained result is related to the fact that the optimum operation

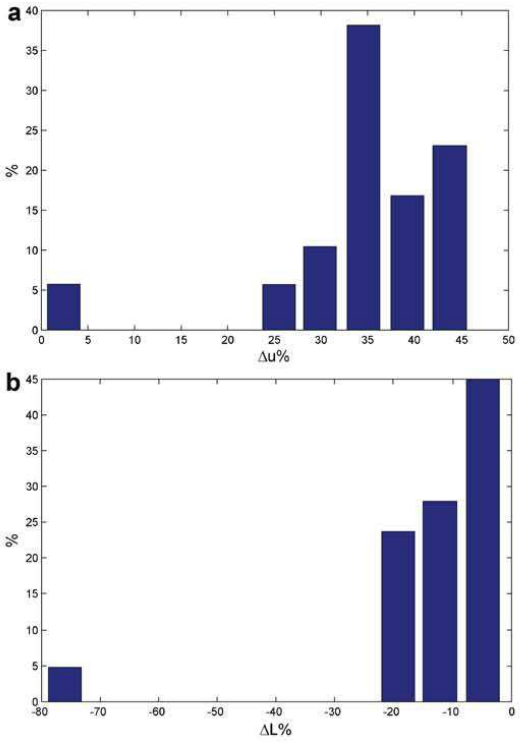


Figure 7. a: Histogram of $\Delta u\%$ for all instances of five consecutive RTO cycles replicated 100 times. b: Histogram of $\Delta L\%$ for all instances of five consecutive RTO cycles replicated 100 times.

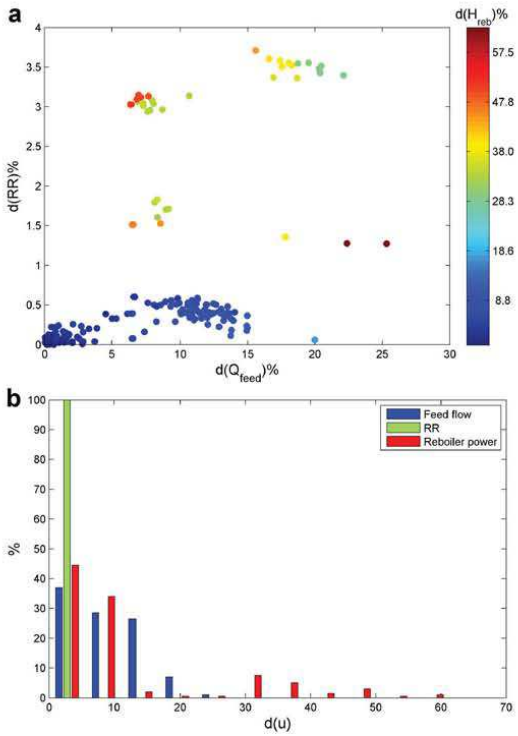


Figure 8. a: Individual values of triplets (d(Q_{feed}), d(RR), d(H_{reb})). d(x): percentage difference of optimised decision variables between two consecutive RTO cycles. Results for 100 repetitions of series of five RTO cycles. b: Histogram for each decision variable of the percentage difference of optimised decision variables between two consecutive RTO cycles. Results for 100 repetitions of series of five RTO cycles.

point is often very close to the proposed process constraint so that constraint violation is very sensitive to modification of the model structure and model parameters.

It is possible to improve RTO performance by altering its structure, as shown in Figure 9. However, since RTO practitioners have to pay attention to many aspects of the implementation, some contradictory findings can make the decision-making process more difficult. This particular example is useful to highlight such difficulty. It can be seen in Table 8 that the set (UA,η) of updated parameters, previously considered the "worst" choice, is the one that shows the lowest probability to cause constraint violation along successive RTO cycles under the assumed conditions. It should be noticed that $\Delta L\%$ values in Table 8 refers only to cycles that do not produce violations.

Constraint violation is a major concern at plant site, since it is not only a matter of profit, but also a matter of process safety. Unfortunately, standard two-step RTO is very vulnerable to such an issue, since most important process constraints are defined in terms of the output values and can only be predicted accurately when the model is perfect and the process data are not affected by measurement noise. One possible approach to avoid frequent constraint violations is to use additive safety margins or back-off from the constraints (Contreras-Dordelly and Marlin, 2000; Kookos, 2003; Govatsmark and Skogestad, 2005), as described in Equation (25). However, this is equivalent to imposing much harder constraints on the process operation than necessary for safe

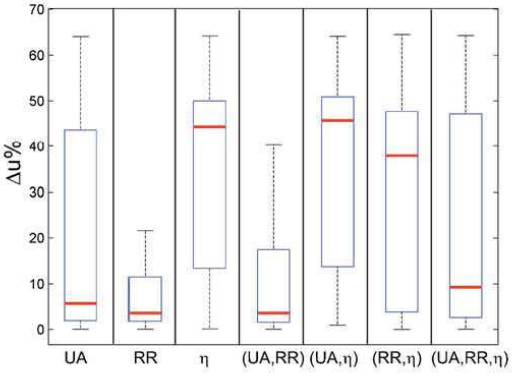


Figure 9. Boxplot of the impact of different choices of updated parameters on the distribution of $\Delta u\%$. The text along the horizontal axis describes the variables updated during the estimation problem. Rectangles contain data between 25° and 75° percentiles. Vertical bars indicate range from 5° to 95° percentiles.

Table 8. Impact of different choices of updated parameters on the median of $\Delta u\%$, $\Delta L\%$ and on the percentage of RTO cycles showing constraint violations

	UA	RR	η	UA,RR	UA, η	RR, η	UA, RR, η
$\Delta u\%$	6	3	45	4	46	39	14
$\Delta L\%$	−4	−1	−15	−2	−15	−12	−11
% violations	28	44	18	34	6	25	15

Results for hybrid PSO–NM Simplex (population = 40; 5 generations) optimisation method.

and profitable operation. As a consequence, profits are expected to decrease to suboptimal values.

$$\mathbf{g}(\mathbf{I},\mathbf{O}) + \boldsymbol{\beta} \leq \mathbf{0} \tag{25}$$

Few works (de Hennin, 1994; Loeblein and Perkins, 1998; Loeblein et al., 1999) care about constraint violations in the context of the two-step RTO. The main problem in this case is to properly design the magnitude and direction of back-off vector β. As a matter of fact, β is dependent, among other factors, on the p.d.f. of noise, $\psi(\varepsilon)$, on the actual operational point, I, and on the specified confidence level (which is related to the conservatism of the back-off). Those factors are better dealt if the optimisation problem is solved in the context of stochastic programming (Kall and Wallace, 1994), since uncertainty is explicitly made part of the optimisation problem. However, stochastic programming is time consuming and presents numerical difficulties if solved as a nonlinear program. Few articles (Zhang et al., 2002; Li et al., 2008; Mesfin and Shuhaimi, 2010) use this approach to RTO problems but under the restrictive assumptions that uncertainty appears linearly in constraint inequalities.

Ethylene Production

Ethylene can be produced through the selective cracking of naphtha or light hydrocarbons (ethane and propane) in tubular reactors placed inside furnaces. The furnace can be divided into two sections: the first one is the radiation section, where the gas fuel is burned, providing energy to the process, and thermal cracking reactions occur; the second one is the convection section, where part of the energy generated during combustion is reutilised and the feedstock is prepared for the cracking reaction. Typical industrial sites contain 10–20 furnaces, which are fed by streams of different compositions in order to meet the desired product blend composition and allow for maximum profit. Therefore, the most important tasks of RTO systems in these processes is the design of the feed conditions (feed flowrates and composition) for each furnace and of the gas fuel flowrates required to meet the optimum operation temperatures.

In order to perform the proposed task, a detailed process model (Jesus, 2011) was implemented in line and incorporated into the framework of a RTO scheme in an industrial site. Since the heat transfer conditions at each distinct furnace vary with time, estimation of heat transfer coefficients is of fundamental importance for proper adaptation of the model responses. However, as the characteristic process time is very short (no more than a few seconds) and the feed composition is subject to significant fluctuations (the feed composition is affected significantly by recycle streams), estimation must also consider the reconciliation of available process data (temperatures, pressures, input and output stream compositions), which are made available with specified sampling frequencies.

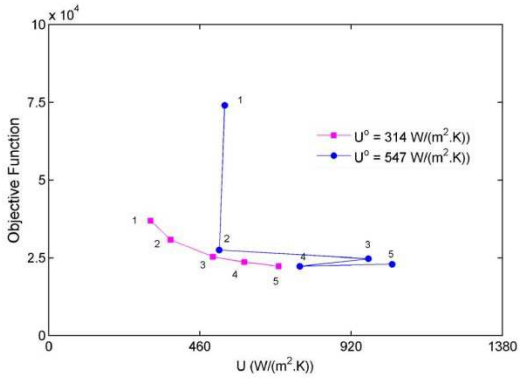


Figure 10. Typical evolution of the objective function when the estimation of the heat transfer coefficient is performed independently (purple squares) and simultaneously (blue circles) with the reconciliation of process data.

As described in the previous sections, the joint reconciliation of process data and estimation of process parameters imposes the implementation of parameter estimation procedures in spaces of very high dimensions. The estimability problem is illustrated in Figure 10, where one can see typical trajectories of the estimated heat transfer coefficients when the parameter estimation is performed independently and simultaneously with the data reconciliation procedure, using real process data. As one can see in Figure 10, the trajectories are subject to significant variations when the data reconciliation procedure of all process inputs are performed simultaneously, while trajectories are smooth and well-behaved when the estimation of the heat transfer coefficient is performed independently. This probably indicates that some of the model parameters are not estimable, either because they are correlated to other parameters and process data or because they do not affect the objection function significantly. Figure 11 reinforces this interpretation, since the estimated value of the heat transfer coefficient starts to oscillate vigorously at plant site when reconciliation of all process inputs is initiated. It seems obvious that the estimation problem must be carefully designed for proper tuning of the RTO performance.

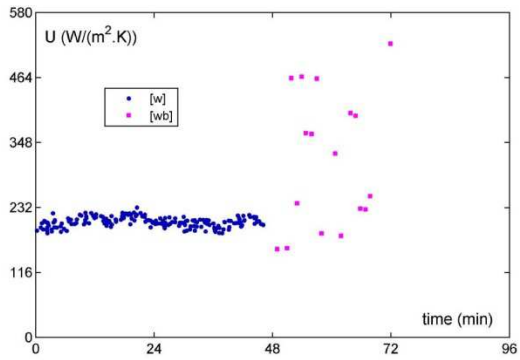


Figure 11. Estimates of the heat transfer coefficient during normal furnace operation before (w) and after (wb) initiation of the reconciliation of all process inputs.

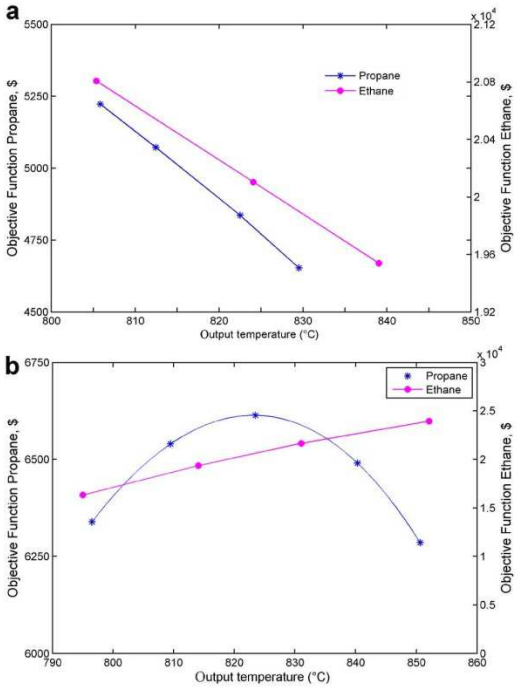


Figure 12. a: Expected profits when each furnace is optimised independently. b: Expected profits when all furnaces are optimised simultaneously.

In this particular problem, the long-term analysis of the Fisher information matrix indicates that the number of estimable parameters changes from three to six, depending of the feed composition and available process data. This can probably encourage the implementation of estimability procedures online, although this is not feasible in most real processes because of the computation time required to build the Fisher information matrix (computation of second derivatives in respect to all pairs of estimated variables) and compute characteristic values and vectors. As observed at plant site, estimability problems were related to the high correlation between certain pairs of variables and low sensitivity of the objective function to the modification of some of the process inputs.

A very interesting and intriguing question is related to Figure 12. Figure 12 shows the profits predicted by the RTO procedure when the furnaces are analysed independently (each furnace is optimised independently from the others) and simultaneously (the optimisation problem is defined in a higher dimension). One can see that the optimum operation conditions (in this case, represented in terms of the desired output temperature) are placed at very distinct operation regions when either pure propane or pure ethane is used as feed. The reason why this strange result is obtained is easy to explain—when the furnaces are optimised independently, the feed streams must be defined a priori in order to meet the desired product blend. Besides, when all furnaces are considered simultaneously, the distillation constraints can be inserted into the optimisation problem in more realistic terms, since all product streams are mixed prior to product purification.

The intriguing problem posed by Figure 12 regards the proper definition of the process envelope. Since the optimisation results change when the process envelope is enlarged to include all furnaces simultaneously, there is no guarantee that the optimum operation conditions will remain the same when additional process sections and equipment are included in the analysis (such as the feed purification and the ethylene polymerisation sections). In other words, since the final objective of the RTO procedure is to provide the maximum profits to the whole corporation, one may be tempted to say that this can only be possible when all aspects of the business are taken into consideration, which is certainly not feasible with sufficiently detailed process models, as described in the previous sections. One should not minimise the fact that prices of intermediate products are subject to significant uncertainties, especially if they are not sold as finished products for other companies, as in this particular case (ethylene is used for production of polyethylene and propylene is used for production of polypropylene). Therefore, it seems reasonable to wonder about the real meaning of the optimisation results that RTO procedures are providing as references for process operation.

FINAL REMARKS

Incomplete information is a silent cause of many problems during RTO implementations. As a matter of fact, many decisions related to the instrumentation network are made prior to RTO design. Even if included in the scope of the RTO project, many decisions related to instrumentation are not commonly taken from rigourous analysis of the RTO performance, as proposed by Fraleigh et al. (2003). Proper choice of sensor technology, sensor placement and measurement redundancy would help the RTO to face the scenarios of disturbances and provide better results in terms of parameter estimability and gross error detection.

The two-step RTO approach is particularly sensitive to model mismatch. In the case of imperfect model structure, optimisation cannot provide the best set of decision variables, no matter whether objective function for model fitting is a pragmatic version of the least-squares procedure or a stricto sensu maximum likelihood function. Another reason for concern is that even if the model is correct, there is no guarantee that the numerical method will find the optimum for both optimisation steps.

Proper selection of the updated parameter set is a much underestimated problem, and the use of updated parameters for process diagnosing is a much overestimated benefit of RTO systems. In huge systems with hundreds of variables eligible for updating, it is unlikely that the use of non-rigourous, experience-related criteria for choosing updated parameters may lead to good results. As a matter of fact, most time updated parameters are just tools that the model fitting procedure uses to manipulate the information and accommodate any disagreement between the model and measurements. This disagreement may be caused by any factor that impairs information acquisition and processing, such as: (i) measurements signals corrupted by noise with unknown error structure; (ii) variation of elements of input I neither measured nor estimated ($\mathbf{var} \not\subset (\mathbf{ms} \cup \mathbf{upd})$); (iii) use of wrong default values for fixed variables; (iv) use of inaccurate process model; (v) violation of maximum likelihood assumptions; (vi) imperfect steady state detection or gross error filtering and (vii) imperfect numerical optimisation method.

In the presence of at least one of the above conditions, estimated parameters will not mimic their real world counterparts. They will instead be the mathematical expression of these influences filtered by the parameter estimation procedure. If the process is complex

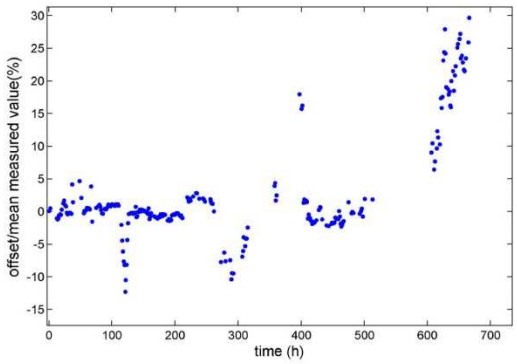


Figure 13. Relative value of estimated offsets for a temperature measurement. Values produced by a real RTO system in a crude distillation unit.

enough, the presence of many items of this list combined with an intuitive choice of updated parameters will pose a serious threat to the production of meaningful variable estimates, ready to be used to diagnose equipments and sensors. Figure 13 illustrates this discussion with data collected in a real industrial site and shows estimated offset values of a temperature measurement made by a RTO system in a crude distillation unit. From the prior knowledge of how temperature sensors work, it would be expected that the estimated values showed some short-term variation due to noise. It is also expected that estimates would present an almost constant non-zero mean value in case of a miscalibrated sensor. However, there is no easy explanation in terms of sensor operation for the sudden changes observed in Figure 13. The variable magnitude and duration of these events are probably caused by many of the reasons discussed, although it is not easy to say which ones. An easier way to explain Figure 13 is to blame the sensor itself. The only problem with this explanation is that this behaviour is not uncommon and too many sensors would have to be considered faulty.

From the point of view of the practitioner, it may be a very difficult task to diagnose an RTO system based on the two-step approach. It is possible that some guidance may be provided by automated analyses produced by specific software tools, but chances are that many conclusions will be very dependent on non-validated assumptions such as: (i) model correctness; (ii) known noise structure and (iii) certainty that all variable input information is either measured or included in the set of updated parameters. In complex cases, it may be virtually impossible for the user to produce higher level diagnostics, due to the overlapping of too many causes of failure.

It is common that RTO projects are very focused on process modelling, paying special attention to the ability to deal with very large and complex sets of highly customisable equipment. The magnitude of these models may give the false impression that only minor problems remain outside the scope of model building. However, many underestimated decisions related to RTO structure may severely impact the system performance. Due to the lack of online and off-line supporting decision tools, the users are requested to use their common sense to make many of decisions:

(a) To choose the set of updated parameters [$\mathbf{upd} = ?$, in Equation (9)].
(b) To choose the set of decision variables [$\mathbf{df} = ?$, in Equation (1)].
(c) To tune the parameter updating objective function by altering values of standard deviations [$Vz = ?$, in Equation (17)].
(d) To decide if the set of optimised decision variables should be implemented.
(e) To customise the objective function and constraints in order to incorporate empirical knowledge ($\mathbf{obj} = ?$, in equation (16), addition of new terms in constraints g, among other decisions).
(f) To provide "good initial estimates" to the optimiser.

Taking all these decisions in a coordinated way may eventually be beyond human abilities. Nevertheless, it is common that RTO commercial systems expect that users perform this task without further automated support based on rigorous analysis. It would be interesting to add some comments on the previous list of user decisions:

(a) Updated parameters should work on the best interest of profit optimisation. Their role should be more mathematical than physical. In a pragmatic way, parameter updating should confer robustness to RTO in face of the previously presented list of factors that impairs information acquisition and processing. In this sense, use of intuition in a high dimension space of parameters may be misleading.
(b) The choice of the set of decision variables should take into account direct correspondence not only with the available degrees of freedom, but also with the expected variability of the results. Any change in this set along operation requires the re-evaluation of all other choices.
(c) If standard deviations are to be used as tuning parameters, it should be supported by an automated optimisation subsystem taking care of this choice from the point of view of robustness of the RTO results. The popular question "Which measurements do you trust more?" cannot be properly answered to fulfill the requirements of good RTO performance if solely supported by experience. On the other hand, if standard deviations are to be used in the context of rigourous maximum likelihood estimation, proper methods of experimental evaluation of the covariance matrix and the p.d.f. of noise have to be employed. Intermediate approaches will lack the virtues of both approaches. The statistical foundation will be lost without improving global performance of RTO.
(d) Ideally, changes in decision variables should always lead to better economic performance. Unfortunately, several factors, as previously discussed, may add useless variability to decision variables. Very few academic articles (Miletic and Marlin, 1996; Zhang et al., 2001) pay attention to this decision about the decision variables: should they be implemented or not? In practice, it is common that empirical rules have to be used by the practitioners, such as the use of a profit improvement threshold that triggers the acceptance. The problem with empirical rules is that they demand other empirical rules (the minimum value of profit improvement) and may be too optimistic (predicted profit is different from actual profit, due to model and parameter mismatches). Validation of the values of the decision variables by the operator is commonly the last step before acceptance. Unfortunately, it is a hard task to inspect a vector of several variables. This visual inspection will never be properly made in the high-dimensional space of variables. Eventually it will become a series of unidimensional bound checks.

(e) It is a common procedure to impose bounds to the maximum variation of updated parameters and decision variables between successive RTO cycles. The intention is to avoid unrealistic values of estimated parameters and unnecessary disturbances to the process. Some "safe margin" may also be added to the constraints in order to decrease the chance of constraint violation. Although these are reasonable aims, the common procedure of figuring out the limits is by empirical guidance. It is also monovariable in its essence, since each single limit is fixed without further concerns over the whole set of variables.

(f) It is a well-known fact that deterministic methods are prone to get trapped in local minimum (Yang, 2010). Convergence is also an issue. Although suboptimal results are a hidden problem, lack of convergence is evident. It is very common that RTO systems demand that the user provide "better" initial estimates if optimisation does not converge. This may be difficult, since the set of initial estimates is usually large and the very concept of a "better" initial estimate is unclear. If "better" means nearer to the optimum, the user cannot give much help, since s/he uses the optimiser precisely because s/he does not know where the optimum lies. The best guess the user can provide is the actual operational point or somewhere in its vicinity. As a matter of fact, from the point of view of the numerical optimisation method, "better initial estimate" does not necessarily mean closer to the optimum. It rather means a point that put the optimiser on the right track to the optimum. This concept implies that the user should figure out the way the optimiser operates in a specific scenario (i.e. curvature, discontinuity), which is not achievable, in most of cases.

As one can see, the two-step RTO approach faces several challenges in real world implementations. Although the general tone of this article is not enthusiastic, the main intention is to give users some insights on common causes of performance degradation that may contribute to a more fruitful interaction with solution providers. Although the two-step approach has several weaknesses, its implementation is far more feasible than the other RTO approaches (Chachuat et al., 2009).

Common RTO systems are too optimistic on the validity of project assumptions. One can argue that many assumptions are very reasonable and that some possible departure does not affect performance in practical terms. It is possible that this can be true, but such a conclusion must be derived from proper analysis instead of made on a priori basis. Thorough investigation has to be performed in order to make clear that actual profit improvement exceeds costs of the project, implementation and operation under uncertainty and maintenance. In common RTO practice, an extensive list of decisions is transferred to the user. The absence of proper diagnostic tools, combined with the multivariable and integrated character of these decisions, creates a huge and unfair task to the user.

In face of the fact that all decisions in an RTO system are interdependent, the only reasonable way to make the necessary choices is to make them simultaneously. In this sense, there are two approaches that should be part of RTO projects and routine operation. They are the 'Àverage Deviation from Optimum" (de Hennin, 1994; Loeblein and Perkins, 1998) and the "Design Cost" (Forbes and Marlin, 1996; Zhang and Forbes, 2000) approaches. These methods provide a better way to integrate several decisions, such as measurement selection, choice of the set updated parameters, design of backoff vector taking into account the mean performance of RTO over several scenarios of disturbances, failures and model mismatch. Since the two-step approach is far from perfect, the only way to make RTO more robust is to carefully consider all its vulnerabilities, to spend time enumerating all operation scenarios and to design proper diagnostic tools for decision support.

CONCLUSIONS

The implementation of optimisation procedures in real time (RTO) for improvement of process performance constitutes a major challenge for those involved with process operation. However, implementation of RTO procedures must necessarily rely on the availability of robust process models, numerical methods and process data. As shown through many examples, however, the real performances of process models and numerical methods and the quality of real process data exert very significant negative effects on the performances of RTO procedures. Some vulnerabilities related to numerical performance, model formulation and quality of process data were described and discussed in the framework of the two-step RTO structure commonly found in commercial software. Since the two-step approach is far from perfect, the only way to make RTO more robust is to carefully consider all the considered vulnerabilities, to spend time enumerating all operation scenarios and to design proper diagnostic tools for decision support.

NOMENCLATURE

df	indexes of elements of I that are degrees of freedom for economic optimisation
f	process model equations
fix	indexes of non-updated elements of Θ
g	process constraints
H_{reb}	power delivered to reboiler (kW)
I	input variables
L	economic objective function
ms	indexes of measured elements in Z
obj	indexes of elements of z included in the objective function of parameter estimation
O	output variables
Q_{feed}	feed flow (kmol/h)
Q_{tp}	molar flow, top product (kmol/s)
RR	reflux ratio
std	indexes of elements of I that remain constant along RTO cycles
u	decision variables
UA	global heat transfer coefficient times the area of top condenser
um	indexes of unmeasured variables in Z
upd	indexes of updated elements in Θ
var	indexes of elements of I that varies along RTO cycles
V	variance-covariance matrix
z	subset of the measured elements in Z
Z	whole set of variables in process model
$\$_{pe}$	propene price (monetary unit/kmol)
$\$_H$	cost of power delivered to reboiler (monetary unit/kW)

Greek Symbols

δ	designates the offset of a variable in the vector of parameters, Θ
$\Delta u\%$	norm of the percentage difference of u
$\Delta L\%$	percentage difference of L
$\boldsymbol{e}$	noise vector

η Murphree tray efficiency
θ subset offset modifiers chosen to be updated
Θ set of offset modifiers
ψ probability density function

ACKNOWLEDGEMENTS

We thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil) for providing scholarships and supporting our work. We also thank Petrobras (Petróleo Brasileiro SA) and Braskem Indústrias Químicas SA for supporting our work.

REFERENCES

Bard, Y., "Nonlinear Parameter Estimation," Academic Press, New York (1974).

Basak, K., K. S. Abhilash, S. Ganguly and D. N. Saraf, "On-Line Optimization of a Crude Distillation Unit with Constraints on Product Properties," Ind. Eng. Chem. Res. 41, 1557–1568 (2002).

Bates, D. M. and D. G. Watts, "Nonlinear Regression Analysis and its Applications," John Wiley and Sons, New York (1988).

Biegler, L. T., I. E. Grossman and A. W. Westerberg, "A Note on Approximation Techniques Used for Process Optimization," Comput. Chem. Eng. 9, 201–206 (1985).

Bjorkman, M. and K. Holmstrom, "Global Optimization Using the DIRECT Algorithm in Matlab," Adv. Model. Optim. 1, 17–37 (1999).

Chachuat, B., B. Srinivasan and D. Bonvin, "Adaptation Strategies for Real-Time Optimization," Comput. Chem. Eng. 33, 1557–1567 (2009).

Chen, C. Y. and B. Joseph, "On-Line Optimization Using a Two-Phase Approach—An Application Study," Ind. Eng. Chem. Res. 26, 1924–1930 (1987).

Contreras-Dordelly, J. L. and T. E. Marlin, "Control Design for Increased Profit," Comput. Chem. Eng. 24, 267–272 (2000).

Cubillos, F. A., G. Acuña and E. L. Lima, "Real-Time Process Optimization Based on Grey-Box Neural Models," Braz. J. Chem. Eng. 3, 433–443 (2007).

Darby, M., M. Nikolaou, J. Jones and D. Nicholson, "RTO—An Overview and Assessment of Current Practice," J. Process Contr. 21, 874–884 (2011).

Datskov, I., G. M. Ostrovsky, L. E. K. Achenie and Y. M. Volin, "Process Optimization Under Uncertainty When There Is Not Enough Process Data at the Operation Stage," Optim. Eng. 8, 249–276 (2006).

de Hennin, S. R., "Structural Decisions in On-line Process Optimization," Ph.D. Thesis, University of London (1994).

Fletcher, R., "Practical Methods of Optimization," Wiley, New York (1987).

Forbes, J. F. and T. E. Marlin, "Design Cost: A Systematic Approach to Technology Selection for Model-Based Real-Time Optimization Systems," Comput. Chem. Eng. 20, 717–734 (1996).

Forbes, J. F., T. E. Marlin and J. F. MacGregor, "Model Adequacy Requirements for Optimizing Plant Operations," Comput. Chem. Eng. 18, 497–510 (1994).

Fraleigh, L., M. Guay and J. F. Forbes, "Sensor Selection for Model-Based Real-Time Optimization: Relating Design of Experiments and Design Cost," J. Process Contr. 13, 667–678 (2003).

Friedman, Y. Z., "Closed-Loop Optimization Update—A Step Closer to Fulfilling the Dream," Hydrocarb. Process 79, 15–16 (2000).

Gattu, G., S. Palavajjhala and D. B. Robertson, "Are Oil Refineries Ready for Non-Linear Control and Optimization?" in: International Symposium on Process Systems Engineering and Control, January, Mumbai (2003).

Golshan, M., M. R. Pishvaie and R. B. Boozarjomehry, "Stochastic and Global Real Time Optimization of Tennessee Eastman Challenge Problem," Eng. Appl. Artif. Intel. 21, 215–228 (2008).

Govatsmark, M. S. and S. Skogestad, "Selection of Controlled Variables and Robust Setpoints," Ind. Eng. Chem. Res. 44, 2207–2217 (2005).

Jesus, N. J. C., "Otimização em Tempo Real em um Processo Industrial de Produção de Etileno," PhD Thesis, PEQ/COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro (in Portuguese) (2011).

Kall, P. and S. W. Wallace, "Stochastic Programming," John Wiley and Sons, Chichester (1994).

Kookos, I. K., "Optimal Operation of Batch Processes under Uncertainty: A Monte Carlo Simulation-Deterministic Optimization Approach," Ind. Eng. Chem. Res. 42, 6815–6822 (2003).

Koppen, M., "No-Free-Lunch Theorems and the Diversity of Algorithms," in Proceedings of the 2004 IEEE Congress on Evolutionary Computation (CEC-2004), 235–241 (2004).

Li, P., H. Arellano-Garcia and G. Wozny, "Chance Constrained Programming Approach to Process Optimization Under Uncertainty," Comput. Chem. Eng. 32, 25–45 (2008).

Loeblein, C. and J. Perkins, "Economic Analysis of Different Structures of On-Line Process Optimization Systems," Comput. Chem. Eng. 22, 1257–1269 (1998).

Loeblein, C., J. D. Perkins, B. Srinivasan and D. Bonvin, "Economic Performance Analysis in the Design of On-Line Batch Optimization Systems," J. Process Contr. 9, 61–78 (1999).

McLean, K. A. P. and K. B. McAuley, "Mathematical Modelling of Chemical Processes—Obtaining the Best Model Predictions and Parameter Estimates Using Identifiability and Estimability Procedures," 90(2), 351–366 (2012).

Mesfin, G. and M. Shuhaimi, "A Chance Constrained Approach for a Gas Processing Plant With Uncertain Feed Conditions," Comput. Chem. Eng. 34, 1256–1267 (2010).

Miletic, I. P. and T. E. Marlin, "Results Analysis for Real-Time Optimization (RTO): Deciding When to Change the Plant Operation," Comput. Chem. Eng. 20, 1077–1082 (1996).

Miletic, I. P. and T. E. Marlin, "Results Diagnosis for Real-Time Process Operations Optimization," Comput. Chem. Eng. 22, 8475–8482 (1998).

Nelder, J. A. and R. Mead, "A Simplex Method for Function Minimization," Comput. J. 7, 308–313 (1965).

Prata, D. M., M. Schwaab, E. L. Lima and J. C. Pinto, "Nonlinear Dynamic Data Reconciliation and Parameter Estimation Through Particle Swarm Optimization: Application for an Industrial Polypropylene Reactor," Chem. Eng. Sci. 64, 3953–3967 (2009).

Roberts, P. D. and T. W. C. Williams, "On an Algorithm for Combined System Optimisation and Parameter Estimation," Automatica 17, 199–209 (1981).

Schwaab, M., E. C. Biscaia, Jr., J. L. Monteiro and J. C. Pinto, "Nonlinear Parameter Estimation through Particle Swarm Optimization," Chem. Eng. Sci. 63, 1542–1552 (2008).

Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," John Wiley and Sons, New York (1989).
Volin, Y. M. and G. M. Ostrovskii, "Process Optimization Under Insufficient Experimental Information in the Phase of Service," Automat. Rem. Control 66, 1195–1211 (2005).
White, D. C., "Online Optimization: What Have We Learned?" Hydrocarb Process 77, 55–59 (1998).
Wolpert, D. W. and W. G. Macready, "No Free Lunch Theorems for Optimization," IEEE Trans. Evol. Comp. 1, 67–82 (1997).
Yang, X.-S., "Engineering Optimization: An Introduction with Metaheuristic Applications," Wiley, USA (2010).
Yip, W. S. and T. E. Marlin, "Designing Plant Experiments for Real-Time Optimization Systems," Control Eng. Pract. 11, 837–845 (2003).
Zhang, Y. and F. Forbes, "Extended Design Cost: A Performance Criterion for Real-Time Optimization Systems," Comput. Chem. Eng. 24, 1829–1841 (2000).
Zhang, Y., D. Monder and J. F. Forbes, "Real-Time Optimization Under Parametric Uncertainty: A Probability Constrained Approach," J. Process Contr. 12, 373–389 (2002).
Zhang, Y., D. Nadler and J. F. Forbes, "Results Analysis for Trust Constrained Real-Time Optimization," J. Process Contr. 11, 329–341 (2001).

Manuscript received February 5, 2012; revised manuscript received June 9, 2012; accepted for publication June 15, 2012.